MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABCRATORY

A.I. Memo 768 August, 1984

ON EDGE DETECTION

V. Torre and T. Poggio

Abstract. Edge detection is the process that attempts to characterize the intensity changes in the image in terms of the physical processes that have originated them. A critical, intermediate goal of edge detection is the detection and characterization of significant intensity changes. This paper discusses this part of the edge detection problem. To characterize the types of intensity changes derivatives of different types, and possibly different scales, are needed. Thus, we consider this part of edge detection as a problem in numerical differentiation.

We show that numerical differentiation of images is an ill-posed problem in the sense of Hadamard. Differentiation needs to be *regularized* by a regularizing filtering operation before differentiation. This shows that this part of edge detection consists of two steps, a *filtering* step and a *differentiation* step. Following this perspective, the paper discusses in detail the following theoretical aspects of edge detection:

(1) The properties of different types of filters—with minimal uncertainty, with a bandpass spectrum, and with limited support—are derived. Minimal uncertainty filters optimize a tradeoff between computational efficiency and regularizing properties.

(2) Relationships among several 2-D differential operators are established. In particular, we characterize the relation between the Laplacian and the second directional derivative along the gradient. Zero-crossings of the Laplacian are not the only features computed in early vision.

(3) Geometrical and topological properties of the zero crossings of differential operators are studied in terms of transversality and Morse theory.

We discuss recent results on the behavior and the information content of zero crossings obtained with filters of different sizes. These results imply a specific order in the sequence of filtering and differentiation operations. Topological properties are preserved by level-crossings. Setting a level in the optimal filtering stage is a threshold operation — which can be implemented in an adaptive way — that preserves all the "nice" geometrical and topological properties of zero crossings.

Finally, some of the existing local edge detector schemes are briefly outlined in the perspective of our theoretical results.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-80-C-0505. Funds for collaborative travel were provided by NATO grant 237.81. Dr. Torre is at the University of Genoa, Italy.

Contents

1. INTRODUCTION

Vision begins with the transformation of a flux of photons into a set of intensity values at an array of sensors. The first step in visual information processing is to obtain a compact description of the raw intensity values. The primitive elements of the initial description should ideally be *complete* in the sense of representing the full information contained in the image, and *meaningful* (that is, capturing significant properties of the three-dimensional surfaces around the viewer). Physical edges are one of the most important properties of objects since they correspond to object boundaries or to changes in surface orientation or material properties (Ballard and Brown, 1982; Binford, 1981, 1982; Brady, 1981; Canny, 1983; Davis, 1975; Hildreth, 1980; Marr and Hildreth, 1980; Pavlidis, 1977; Rosenfeld and Kak, 1976).

Three-dimensional edges are often mapped by the imaging process into critical points of the two-dimensional intensity profile formed in the eye or in a camera. The ultimate goal of edge detection *is the characterization of intensity changes in the image in terms of the physical processes that originated them.* For instance, a shadow may be distinguished from an occluding boundary and material properties may be identified from the associated intensity changes.[1] A traditional belief in computational vision—that we fully share—is that this goal cannot be reached in a single step. At least two separate stages are required. First, one needs to characterize the intensity changes in the image. Second, one uses this representation, combined with high-level knowledge, to make assertions about the 3-D surfaces and their properties.

The first part of edge detection then, requires the evaluation of derivatives of the image intensity. To characterize the types of intensity changes, derivatives of different type and order may be needed, possibly at different scales. The first part of edge detection is thus a problem in numerical differentiation. In this paper, we will consider only this first stage of edge detection as the process that attempts to detect, localize and characterize *local edges*, the sharp changes in intensity that are natural primitives for later processing. We will not consider here the second stage of edge detection that includes processes such as boundary detection, segmentation, region growing and groupings of local edges (that group local edge elements into structures better suited for the interpretation of image data in terms of the underlying physical processes).

In this paper, we begin by analyzing the problem of differentiating a sampled image. We show that differentiation is an ill-posed problem (in the sense of Hadamard). Well-posedness and numerical stability of the differentiational step requires the regularization of the image intensities by a regularizing filtering operation preceding differentiation. This argument represents a novel and rigorous justification of the basic sequence of filtering and differentiation that can be recognized in all existing local edge detector schemes. We then examine in detail the filtering and the differentiation stage. We continue our analysis by characterizing properties of the critical points of the differentiation operation.

Our main practical conclusions in this paper are (a) that Gaussian filtering, although not optimal under all conditions, is near-optimal, and computationally convenient; (b) the choice between rotationally invariant operators (rotational filters and rotational invariant differential RID operators, as the Laplacian or the second derivative in the direction of the gradient) or directional (directional filter and directional differential DD operators) operators (such as directional derivatives) depends on the subsequent information processing task. RID operators ensure closed edge contours, that are not provided in general by DD operators.

We now outline the organization of this paper in more detail.

[1] The use of color information—which we will not discuss in this paper—is a natural extension within this framework.

1.1. Organization of the paper

In this paper, we consider *edge detection* as the process of computing derivatives in the two-dimensional intensity image. In Section 2, we show that the problem of differentiation of a sampled image is ill-posed. We prove that filtering of the image prior to differentiation is necessary for regularizing the problem and make it well-posed. The filtering step is analyzed in Section 3. Filters with minimal uncertainty (Hermite and Gabor functions), with bandpass properties (sinc and prolate functions) and others that are support-limited are reviewed. Filter with minimal uncertainty tend to optimize the trade off between band-limited characteristics (required for a correct sampling and for "regularizing" the differential operation) and computational efficiency.

Section 4 is devoted to the differential stage. We consider separately the second order RID and DD operators and analyze their main properties. The main focus is on the localization of the zeros of the Laplacian ∇^2, the second derivative along the gradient $\frac{\partial^2}{\partial n^2}$ and the usual second order partial derivatives. Section 5 considers the geometrical structure of the contours formed by edge detectors and in particular their closure property. For this purpose, we use elementary tools from Morse and Thom theories. The problem of the geometry of contours across different spatial scales—where scale is parameterized by the size of the filter—is considered in Section 6. A comparison of the results of our study with several previously proposed edge detectors is given in Section 7, and a discussion of the "best" filtering and differential steps is given in the final section.

2. COMPUTING DERIVATIVES OF IMAGES

In this chapter, we consider the problem of computing (spatial or temporal) derivatives of sampled intensity images. Our main result is a rigorous justification of filtering before differentiation in terms of the theory of regularization. Our approach also clarifies the issue of the optimal filter for edge detection. In practice, it justifies the use of suitable derivatives of gaussian-like filters in edge detection (for linear differential operators).

In the first section, we discuss the ill-posed nature of differentiation, which is equivalent to its lack of robustness against noise in the input data. In section 2.2, we review the main techniques for transforming differentiation into a well-posed problem. Section 2.3 shows that numerical differentiation can be regularized via previous convolution of the image data with an appropriate filter. In section 2.4, we consider the application of two of the general regularization techniques, and show that they lead to spline interpolation and to spline approximation respectively (prior to the differentiation stage). In these methods, regularized differentiation is thus performed by convolving the data with an appropriate derivative of the regularizing filter. In some situations, however, it may be more convenient in practice to first filter the data and then differentiate the results. We consider some implications of this situation in Appendix 1. The problem of sampling appropriately the image prior to filtering and differentiation is discussed in Appendix 2. Interpolation, approximation and differentiation are discussed in Appendix 4.

2.1. Ill-posed nature of differentiation

In machine vision, as well as in most numerical problems, the data are noisy. Noise in the phototransduction process is ultimately unavoidable. Sensor noise arises at least in

part from quantum fluctuations in the number of absorbed photons per sensor and unit time. This represents a fundamental limitation for real time imagery when integration time and size of the sensors are limited by the need of high temporal and spatial resolution. It is critically important, therefore, that the results of numerical operations performed on the data are not too sensitive to noise. It is well known that differentiation is not robust against noise. Even a small amount of noise may disrupt differentiation. Let us consider a function $f(x)$ and $\hat{f}(x) = f(x) + \epsilon \sin \omega x$. $f(x)$ may be close to $\hat{f}(x)$ according to standard norms ($L^2, L^\infty \ldots$), provided ϵ is sufficiently small. On the other hand, $f'(x)$ may be quite different from $\hat{f}'(x)$ if ω is large.

In the beginning of this century, Hadamard (1923) defined a mathematical problem to be well-posed if its solution

(a) exists

(b) is unique

(c) depends continuously on the initial data (this is equivalent to saying that the solution is robust against noise).

Most of the problems of classical physics are well-posed in this sense, and Hadamard argued that meaningful physical problems had to be well-posed.

Now differentiation of the function $f(x)$ is a typical ill-posed problem, since it can be seen as the solution to the inverse problem

$$g(x) = Af(x) \qquad [2.1]$$

where $Af(x)$ is the integral operator

$$\int_{-\infty}^{x} f(\tilde{x})d\tilde{x} = \int_{-\infty}^{\infty} h(x - \tilde{x})f(\tilde{x})d\tilde{x} \qquad [2.2]$$

where h is the step function. It is well known that inverse linear problems in which $g(x)$ and $f(x)$ belong to Hilbert space are ill-posed (Tikhonov and Arsenin, 1977; Bertero, 1981).

2.2. Regularization techniques

Rigorous methods for transforming ill-posed problems into well-posed problems have been developed over the past years (see especially Tikhonov, 1963; Tikhonov and Arsenin, 1977; and Nashed, 1974, 1976). Regularization of the ill-posed problem of finding z from the data y, such that $Az = y$ requires the choice of suitable norms $\|\cdot\|$, (usually quadratic) and of a stabilizing functional $\|Pz\|$. The choice of the stabilizing functional and of the norms is dictated by mathematical considerations, and most critically, by an analysis of the physical constraints on the problem. There are three main methods of standard regularization (Bertero, 1981):

(1) Among z that satisfy $\|Pz\| \leq C_1$ (where C_1 is a constant) find z that minimizes

$$\|Az - y\| \qquad [2.3]$$

(2) Among z that satisfy $\|Az - y\| \leq C_2$ find z that minimizes

$$\|Pz\| \qquad [2.4]$$

and (3) Find z that minimizes

$$\|Az - y\|^2 + \lambda\|Pz\|^2, \qquad [2.5]$$

where λ is a regularization parameter.

The first method consists of finding the function z that satisfies the constraint $\|Pz\| < C_1$, and best approximates the data. The second method computes the function z that is sufficiently close to the data and is most "regular". In the third method, the regularization parameter λ controls the compromise between the degree of regularization of the solution and its closeness to the data.

Differentiation can also be regularized using the stabilizing operators introduced by Tikhonov (Tikhonov and Arsenin, 1977; Bertero, 1981). In the case of differentiation these operators are equivalent to filtering the data with low-pass filters of the kind we will discuss in chapter 3.

In the next section, we show how to use method 2 and 3 directly for solving the ill-posed problem of numerical differentiation. In section 2.4, we will consider a wide class of regularizing filters that correspond to Tikhonov stabilizing operators and can be used to make numerical differentiation well-posed.

2.3. Regularizing differentiation with interpolating and approximating splines

Poggio, Voorhees and Yuille (1984) have recently applied the second and the third regularizing methods to the problem of edge detection. Following Schoenberg (1964) and Reinsch (1962), they chose for P the simplest form of Tikhonov's stabilizing functionals with $P = \frac{d^2}{dx^2}$ and the usual L_2 norm. This choice corresponds to an "a priori" constraint of smoothness on the intensity function. Its physical justification is that the noiseless image has to be smooth in the sense that all its derivatives must exist and be bounded because the image is band-limited by the optics. Physically, this constraint of smoothness allows us to eliminate effectively the noise that creeps in after or during the sampling and transduction process, and makes the operation of differentiation unstable and ill-posed. This is, of course, not the only stabilizing functional for this problem, as we will see in the next section, but it is probably the simplest one.

Let us now consider in more detail for the second and third regularization methods. Consider a function $f(x)$ defined in $[a, b]$ and be $\Delta = a \leq x_o < x_1 \ldots x_n = b$ a mesh of distinct points, and

$$f_x = f(x_k) \qquad [2.6]$$

the values of $f(x)$ at x_k. Given the sample points of f_k, the problem of computing the numerical derivative f'_k at x_k is ill-posed. The second regularizing method leads (usig the stabilizing operator $P = \frac{d^2}{dx^2}$ and the L_2 norm) to the search of a function $S(x)$ such that

(a)

$$S(x_k) = f_k \qquad k = 1, \ldots, n \qquad [2.7]$$

and (b) $\|PS(x)\|$ is minimized. The stabilizing functional P is

$$\int_a^b |S''(x)|^2 dx \qquad [2.8]$$

The solution to this problem is given by the cubic spline $S_\Delta(x)$ which interpolates $f(x)$ in Δ (Ahlberg, Nilson & Walsh, 1967). As a consequence, the numerical derivative f'_k will be the value of $S'_\Delta(x)$ in x_k. For equidistant points the following equation holds

$$f'_k = \frac{3}{h}\{\alpha(1)(f_{k+1} - f_{k-1}) - \alpha^2(1)(f_{k+2} - f_{k-2}) + \alpha^3(1)(f_{k+3} - f_{k-3}) \ldots\}, \qquad [2.9]$$

where h is the sampling period, and

$$\alpha(x) = \frac{x}{2} - \sqrt{\left(\frac{x}{2}\right)^2 - 1} \qquad [2.9]$$

that is

$$f'_k = \frac{1}{h}[.804(f_{k+1} - f_{k-1}) - .215(f_{k+2} - f_{k-2}) + .0577(f_{k+3} - f_{k-3}) \ldots] \qquad [2.10]$$

Poggio et al. (1984) have obtained the following theorem which is a reformulation of results due to Schoenberg (1946, 1964):

Theorem:*The cubic spline interpolating the data points assumed on a regular lattice and satisfying the second regularizing method with $P = \frac{d^2}{dx^2}$ can be obtained by convolving the data points with the cubic spline filter, which corresponds to the L^4 function of Schoeberg (1946).*

Numerical differentiation, therefore, can be regularized for exact data on a regular grid by convolving the data points with the first derivative of the L^4 filter given by Schoenberg, which is a cubic spline.

In the case of non-exact data which is the most natural situation, the third regularizing method has to be used leading to the problem of finding $S(x)$ such that

$$\sum_{k=1}^{n}(f_k - S(x_k))^2 + \lambda \int |S''(x)|^2 dx \qquad [2.11]$$

is minimum. Both Schoenberg (1964) and Reinsch (1967) proved that approximating cubic splines are the solution to this variational problem. Poggio et al. (1984) have proved the following result:

Theorem:*The solution to the variational problem [2.11] in the case of inexact data on a regular grid (and appropriate boundary conditions), can be obtained (a) by convolving the data with a filter, (b) which is a cubic spline, and (c) which is very similar to a gaussian.*

This implies that regularized differentiation of image data can be performed by convolving the data with the first derivative of a cubic spline filter, which is very close to the gaussian, as shown in figure 1.

This result probably is the simplest and most rigorous proof that a gaussian-like filter represents the correct operation to be performed before differentiation for edge detection. We refer to the paper by Poggio et al. (1984) for a detailed proof of this result and for a comparison between the optimal filter and the gaussian. Poggio et al. (1984) also analyze the role of the regularizing parameter λ, its connection to the optimal scale of the filter, and discuss methods for finding the optimal λ.

2.4. Regularizing filters

In the previous section we have seen that differentiation can be regarded as the inverse problem of the integral equation

$$g(x) = \int_{-\infty}^{x} f(\tilde{x})d\tilde{x} \qquad [2.12]$$

where $f(x)$ must be recovered from the knowledge of the data $g(x)$, which is usually given only on a discrete lattice. This problem is ill-posed, and can be regularized by the

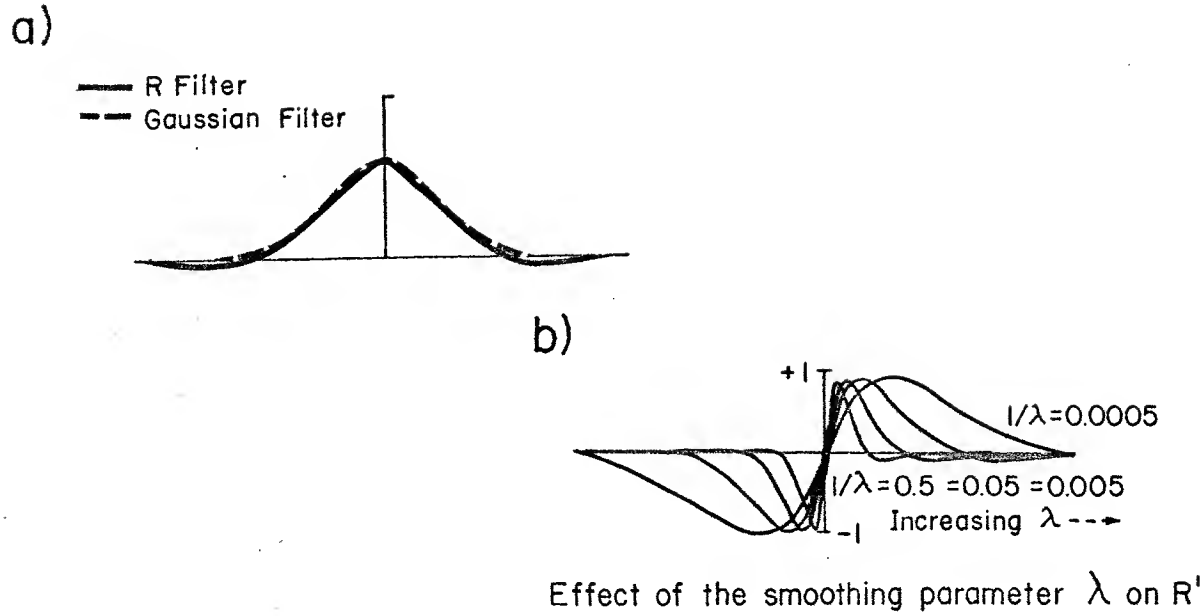

Figure 1 *a) The convolution filter obtained by regularizing the ill-posed problem of edge detection with method (III) (see Poggio et al., 1984). It is a cubic spline (solid line), very similar to a gaussian (dotted line). b) The first derivative of the filter for different values of the regularizing parameter λ, which effectively controls the scale of the filter (from Poggio and Torre, 1984).*

regularizing methods previously mentioned. Furthermore, Tikhonov and Arsenin (1977, see also Bertero, 1981) have proved that it is in general possible to regularize inverse problems by using Tikhonov's stabilizing operators. For equations of the convolution type as equation [2.12], the stabilizing operators correspond to convolving $g(x)$ with a filter $F(x,\alpha)$, (where $\alpha \geq 0$ is a parameter) whose Fourier transform $\tilde{F}(\omega,\alpha)$ satisfies the following conditions:

(C1) $\tilde{F}(\omega,\alpha)$ is bounded for $\alpha \geq 0$ and all ω.

(C2) $\tilde{F}(\omega,\alpha)$ is an even function with respect to ω, and it belongs to $L_2(-\infty,+\infty)$,

(C3) $\tilde{F}(\omega,\alpha)j\omega$ belongs to $L_2(-\infty,+\infty)$.

(C4) For every $\alpha > 0$ it holds $\lim_{|\omega|\mapsto\infty} \tilde{F}(\omega,\alpha) = 0$.

(C5) $\tilde{F}(\omega,\alpha) \mapsto 1$ as $\alpha \mapsto 0$ and $\tilde{F}(\omega,0) = 1$.

This regularizing filter is equivalent to a smooth low pass filter. In the next chapter, we will discuss three different classes of low pass filters that have been used for edge detection. The first two of them fully satisfy the previous conditions (C1-C5), and are therefore regularizing filters in Tikhonov's sense. As a final remark, it is interesting to notice that this regularizing filters usually correspond to the solution of variational principles of the type provided by the third regularization method with an appropriate stabilizer P (compare Tikhonov and Arsenin, 1977, page 121).

3. FILTERING

In this section, we will make some preliminary observations on filtering and then, we will review three kinds of low pass filters, which have been used in machine vision for edge detection. We will consider bandpass filter in section 3.1, support-limited filters in section

3.2 and minimal uncertainty filters in section 3.3. Our conclusion is that bandpass filter as well as minimal uncertainty filters are good regularizing operators for differentiation in the sense of Tikhonov, while support-limited filters are only marginally useful.

As in the study of functions in analysis, many properties of intensity changes can be characterized in terms of zeros of appropriate derivatives. For instance, one-dimensional *step edges* in intensity correspond to extrema of the first derivative, whereas *roof edges* correspond to zeros in the first derivative. The main goal of the filtering and differentiation stage in edge detection is to produce a representation of zeros and extrema. Interestingly, the type of derivative — whether directional or rotationally invariant — and the type of representation — whether zeros or extrema — dictate some general properties of the filter to be used. We will now briefly discuss these two points.

The first point is obvious: directional derivatives require one-dimensional filters properly oriented along the chosen direction; when rotationally invariant operators are used, the filter f is a function of the radial coordinate ρ.

We restrict ourselves to examine linear, space invariant filters. Since isotropy can be assumed, the shape of filters, when viewed one-dimensionally, is an even or odd function. Let us now consider the implications of this for the case of step intensity edges. Because of the arguments developed in the previous section, we detect intensity edges from the zeros of a suitable derivative of the filtered intensity profile (i.e. its critical points). If the shape of the step edge to be detected is $S(x)$, defined as

$$S(x)\begin{cases} = 1 & x > 0 \\ = 0 & x \leq 0 \end{cases},$$

then the output $g(x)$ of the convolution $f(x) * S(x)$ where $f(x)$ is the filter, will be

$$g(x) = F(x) - F(-\infty), \qquad [3.1]$$

with $F(x)$ the integral primitive of $f(x)$. Therefore:

- The extrema of $g(x)$ correspond to the zeros of $f(x)$.
- The zero-crossings of $\frac{d^2}{dx^2} g(x)$ correspond to the extrema of $f(x)$.

Three consequences can be derived from these observations:

1. If we are interested in the extrema of the output $g(x)$, and if we want to have an extremum located at the position of the edge, then $f(x)$ must be an *odd* function.

2. If we are interested in the zero-crossings of $\frac{d^2}{dx^2} g(x)$, and if we want to have a zero-crossing located at the position of the edge, then $f(x)$ must be an *even* function.

3. If we are interested in the extrema or zero-crossings, and if $f(x)$ has many zero-crossings, we will have many secondary extrema or zero-crossings. To avoid false edge detection, $f(x)$ should have the least number of zero-crossings, and the optimal situation would then be:

- If $f(x)$ is odd, then $f(x)$ has only one zero.
- If $f(x)$ is even, then $f(x)$ has no zero.

3.1. Band-limited filters.

Band-limited filters are an obvious choice for regularizing differentiation, since the simplest way to avoid harmful noise is to filter out high frequencies that are amplified by differentiation. Linear and circular prolate functions constitute an especially interesting class of band-limited filters (Frieden, 1971); Landau and Pollack, 1961). Linear prolate functions $\psi_n(x)$ are defined by the relation

$$\int_{-x_o}^{+x_o} \psi_n(x) e^{j\omega x} dx = \frac{i^n 2\pi \lambda_n x_o}{\Omega} \psi_n\left(\frac{\omega x_o}{\Omega}\right), \qquad [3.2]$$

where λ_n are called "linear prolate eigenvalues". From [3.2], we see that $\psi_n(x)$ depends on two parameters, x_o and Ω, whose significance will be seen later. The value of λ_n is a function of $c = x_o\Omega$ and may be written as $\lambda_n(x, c)$. $\psi_n(x)$ depends on c. The main properties of $\psi_n(x)$ are

(1) $\psi_n(x)$ are band-limited.

(2) $\psi_n(x)$ are orthogonal on both the interval $[-x_0, x_0]$ and $[-\infty, +\infty]$, with

$$\begin{aligned} \int_{-x_o}^{+x_o} \psi_n(x)\psi_m(x)dx &= \lambda_n \delta_{nm} \\ \int_{-\infty}^{+\infty} \psi_n(x)\psi_m(x)dx &= \delta_{nm} \end{aligned} \qquad [3.3]$$

(3) $\psi_n(x)$ form a complete set of functions of the space of band-limited functions whose Fourier transform $F(\omega)$ is $F(\omega) = 0$ for $|\omega| \geq \Omega$.

In the defining expression [3.2] of $\psi_n(x)$, there are three constants, Ω, λ_n and x_o. From [3.3], we see that Ω is the cutoff frequency and from [3.3], x_o is half the length of the finite interval over which linear prolate functions $\psi_n(x)$ are orthogonal. From [3.3], we also see that λ_n represents the fraction of energy of $\psi_n(x)$ within $|x| \leq x_o$. The dependence of λ_n on $c = x_o\Omega$ is shown in Figure 4.3 of Frieden (1971). Therefore, once we have chosen Ω, we can find c, and consequently x_o such that the energy of $\psi_n(x)$ is almost completely contained in $|x| \leq x_o$.

Linear prolate functions have the nice property that the band-limited function with cutoff frequency Ω and maximal energy concentrated in $[-x_o, x_o]$ is $\psi_o(x)$ with $c = \Omega x_o$. Similarly, the odd band-limited function with cutoff frequency Ω that has maximal energy concentrated in $[-x_o, x_o]$ is $\psi_1(x)$ with $c = \Omega \cdot x_o$. Linear prolate functions are also useful for solving the inverse problem; that is, the strictly support-limited function in $[-x_o, x_o]$ that has maximally concentrated frequencies in $-[\Omega, \Omega]$ is

$$f = D_{x_o}\psi_o(x, c) \quad c = x_0 \cdot \Omega, \qquad [3.4]$$

where D_{x_0} is the operator defined as

$$D_{x_0} f(x) = \begin{cases} = f(x) & |x| \leq x_o \\ = 0 & |x| > x_o \end{cases}, \qquad [3.5]$$

These results show clearly the difference between $\psi_o(x)$ and sinc(x). They are both band-limited, but $\psi_o(x)$ falls off more rapidly than sinc(x) (see Fig. 2 of Landau and Pollack 1961). On the other hand, the strictly support-limited function, which has the minimal spread of frequencies, is not a Haar function or a Difference Of Boxes filter (see later) but is $D_{x_o}\psi_o(x)$.

Oscillations in the filter may produce ringing phenomena in the edge detection process. To reduce these phenomena, it is necessary to have maximal energy of $\psi_o(x)$ (or $\psi_1(x)$) concentrated in the main lobe. With a value of c equal to 7, more than 99 per cent of the energy of $\psi_o(x)$ and $\psi_1(x)$ is concentrated in $[-x_o, x_o]$.

It is immediate to verify that bandlimited filters satisfy all conditions (see section 3) of Tikhonov in order to regularize differentiation.

If we are interested in rotationally invariant two-dimensional filters that are band-limited, we can simply take even linear prolate functions $\psi_n(r)$, $n = 0, 2, 4\ldots$ and substitute x with

$\sqrt{x^2+y^2}=\rho$. Now $\psi_n(\rho)$ is a band-limited function, but does not have the two-dimensional analog of properties (1)- (3). These properties are satisfied by the circular prolate functions $\Psi_n(\rho)$, defined by relation:

$$\int_0^{\rho_o} J_o(\omega\rho)\Psi_n(\rho)\rho d\rho = (-1)^n \frac{\rho_o}{\Omega}\sqrt{\lambda_n}\Psi_n\left(\frac{\omega\rho_o}{\Omega}\right), \qquad [3.6]$$

where J_o is the Bessel function of order zero.

3.2. Support-limited filters

All real filters have a finite extension and are support-limited. Computational efficiency requires that the support of a filter is as compact as possible. Therefore it is interesting to investigate the properties of filters with strictly limited support. The simplest even filter with a strictly limited support and with unitary energy is

$$f(x)\begin{cases} = 1/\sqrt{2D} & |x| \le D \\ = 0 & |x| > D \end{cases},$$

whose Fourier transform $F(\omega)$ is

$$F(\omega) = \sqrt{\frac{2}{D}}\ \frac{\sin\omega D}{\omega}. \qquad [3.7]$$

In two dimensions, we have

$$f(\rho) = \begin{cases} 1/\sqrt{\pi\rho^2} & \rho \le \rho_o \\ 0 & \rho > \rho_o \end{cases},$$

whose Fourier transform is

$$F(\omega) = \frac{1}{\sqrt{\pi}}\frac{J_1(\omega\rho_o)}{\omega}. \qquad [3.8]$$

This kind of filtering represents the well-known "blurring" of the image through a circular aperture of radius ρ_o.

It is important to observe that this class of support-limited filters fails to satisfy, in a strict sense, the five conditions of section 2.4. In particular, condition (3) ($\tilde{F}(\omega,\alpha)j\omega$ belongs to $L_2(-\infty,+\infty)$ is not satisfied (because differentiation introduces back high frequencies in the same amount as they are removed by this type of filtering). Thus, support limited filters are not good regularizing filters in the sense of Tikhonov. Nonetheless, this class of filters can be still considered as regularizing operators in a weak sense.

If we are interested in odd filters, the simplest support-limited filter is

$$f(x)\begin{cases} = 0 & |x| > D \\ = \frac{1}{\sqrt{2D}} & 0 < x < D \\ = -\frac{1}{\sqrt{2D}} & -D < x < 0 \end{cases}, \qquad [3.8a]$$

whose Fourier transform is

$$F(\omega) = \frac{1}{j\omega}\sqrt{\frac{2}{D}}(1-\cos\omega D). \qquad [3.8b]$$

This filter has already been proposed by Herskovitz and Binford (1970) and is commonly called DOB (Difference of Boxes). It is also a Haar function (see Fig.19 of Harmuth 1972 and page 399 of Kolmogorov and Fomine, 1974). The system of Haar functions is complete and constitutes a basis for all square integrable functions on a bounded interval. This property may have some relevance in the context of image processing with Haar functions. Support-limited filters that are even functions can be easily extended in two dimensions by a simple rotation around the origin. A complete set of support-limited functions in two dimensions, which can be used as filters, is the Haar system with two variables (see Harmuth 1972). The Haar function of equation [3.8] has the nice property of being the optimal support-limited filter that maximizes the signal-to-noise ratio for an ideal step edge, $S(x)$. It is easy to see that spatial spread of $f(x)$ favors the signal-to-noise ratio, while spatial concentration favors localization, for instance of zero-crossings. This can be seen as another formulation of the uncertainty relation (Canny, 1983).

3.3. Filters with minimal uncertainty

In the two previous sections, we analyzed band-limited and support-limited filters. Band-limited filters have theoretically infinite support. A drawback of support-limited filters is that they are regularizing only in a weak sense. It is natural then to try to find an optimal compromise between these two types of filters. A measure of the spread of a function $f \in L^2(\Re)$ in the space and frequency domain is the uncertainty ΔU, defined as:

$$\Delta U = \Omega X, \tag{3.9}$$

where

$$X^2 = \frac{\int_{-\infty}^{+\infty}(x-\bar{x})^2 f^2(x)dx}{\int_{-\infty}^{+\infty} f^2(x)dx} \tag{3.10}$$

$$\bar{x} = \int_{-\infty}^{+\infty} x f^2(x)dx \tag{3.11}$$

$$\Omega^2 = \frac{\int_{-\infty}^{+\infty}(\omega-\bar{\omega})^2|F(\omega)|^2 d\omega}{\int_{-\infty}^{+\infty}|F(\omega)|^2 d\omega}. \tag{3.12}$$

$F(\omega)$ is the Fourier transform of $f(x)$ and

$$\bar{\omega} = \int_{-\infty}^{+\infty} \omega|F(\omega)|^2 d\omega. \tag{3.13}$$

Notice that Ω^2 is proportional to the density of of zero-crossings for Gaussian white noise (Papoulis, 1962; Papoulis, 1965, p. 487). It is well-known that the Gaussian function $e^{\frac{-x^2}{\sigma^2}}$ is the real function $f \in L^2(\Re)$ that minimizes the uncertainty ΔU. On these grounds it has been proposed by Marr & Hildreth (1980) as the optimal filter. The uncertainty of an even or an odd function $f \in L^2(\Re)$ can be easily computed if its representation in terms of Hermite functions is known; that is, if we know the set of c_n such that:

$$f(x) = \sum_{n=0}^{+\infty} c_n \varphi_n(x), \tag{3.14}$$

where

$$\varphi_n(x) = e^{\frac{-x^2}{2}} H_n(x) \qquad [3.15]$$

$H_n(x)$ is the Hermite polynomial of order n. The uncertainty of $\varphi_n(x)$ is simply $n + \frac{1}{2}$. If $f(x)$ is an even function, then

$$f(x) = \sum_{k=0}^{+\infty} c_{2k}\varphi_{2k}(x), \qquad [3.16]$$

and the uncertainty ΔU is given by

$$\begin{aligned} \Delta U &= \sqrt{A^2 - B^2} \\ A &= \frac{\sum_{k=0}^{+\infty}(2k + \frac{1}{2})c_{2k}^2}{\|f\|^2} \\ B &= \frac{\sum_{k=0}^{+\infty} c_{2k+2}c_{2k}\sqrt{(2k+2)(2k+1)}}{\|f\|^2}. \end{aligned} \qquad [3.17]$$

If $f(x)$ is odd,

$$f(x) = \sum_{h=0}^{+\infty} c_{2h+1}\varphi_{2h+1}(x), \qquad [3.18]$$

and the uncertainty ΔU is given by

$$\begin{aligned} \Delta U &= \sqrt{A^2 - B^2} \\ A &= \frac{\sum_{h=0}^{+\infty}(2h + 1 + \frac{1}{2})c_{2h+1}^2}{\|f\|^2} \\ B &= \frac{\sum_{h=0}^{+\infty} c_{2h+3}c_{2h+1}\sqrt{(2h+3)(2h+2)}}{\|f\|^2} \end{aligned} \qquad [3.19]$$

Equations [3.16]–[3.19] follow from properties of Hermite functions. We can easily see that the uncertainty of Hermite functions $\varphi_n(x)$ increases with n as the number of zero-crossings of $\varphi_n(x)$ increases. From these observations, we see that good filters will be composed by Hermite functions with low n. From the point of view of uncertainty, the optimal even filter is $e^{\frac{-x^2}{\sigma^2}}$, and the optimal odd filter is $xe^{\frac{-x^2}{\sigma^2}}$ (the two-dimensional case has been treated by Daugman, 1984a). Another class of functions with small uncertainty consists of Gabor functions

$$\phi_g(x) = e^{\frac{-(x-x_0)^2}{\sigma^2}} e^{j(2\pi f_0 x + \phi_0)}. \qquad [3.20]$$

They are complex functions of a real variable and have uncertainty ΔU equal to $\frac{1}{2}$. However, the real and imaginary part of $\phi_g(x)$ do not have minimal uncertainty. The only real function with uncertainty equal to $\frac{1}{2}$ is the Gaussian.

Filters with minimal uncertainty, as well as bandlimited filters, satisfy the conditions of section 3 in order to regularize differentiation.

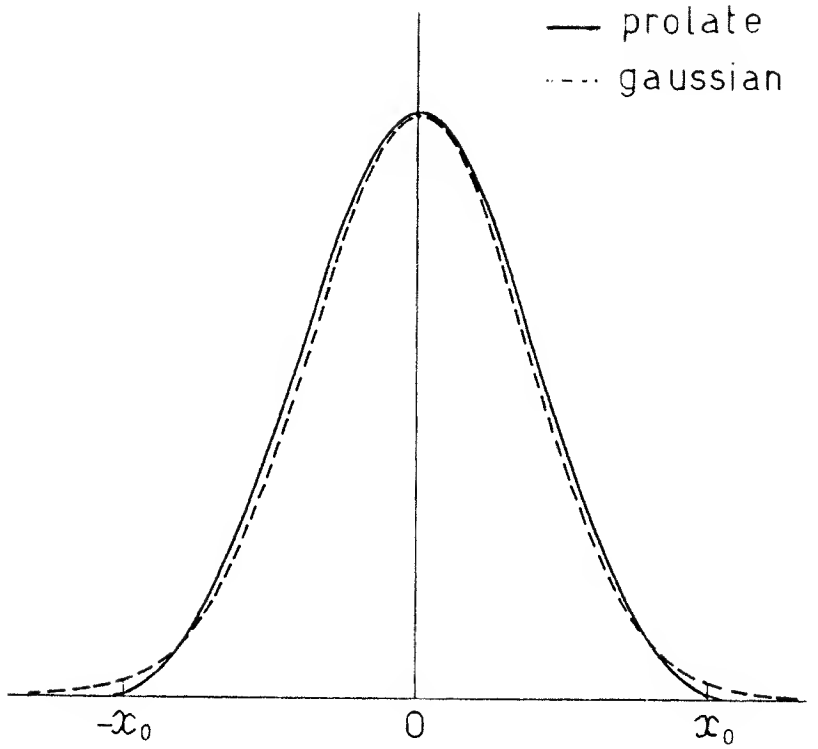

Figure 2 *Comparison between the gaussian and a prolate function. See text.*

3.3.1. Relation between prolate and Hermite functions

The essential difference between prolate and Hermite functions is that the former are band-limited and fall as $\frac{1}{x}$, while the latter fall off faster—somewhat too fast to be band-limited. It has been shown, however, that a crude approximation of $\phi_n(x)$, when c is large (see Slepian 1965) is

$$\psi_n(x) \simeq D_n(x\sqrt{2c}), \qquad [3.21]$$

where D_n is a Weber parabolic cylinder function. Now

$$D_n(x\sqrt{2c}) = e^{\frac{-cx^2}{2}} H_n(x\sqrt{c}) \simeq \psi_n(x), \qquad [3.22]$$

where $H_n(x)$ are the usual Hermite polynomials. This approximation fails for large x, where $\psi_n(x)$ falls off as $\frac{1}{x}$ and $D_n(x)$ as a Gaussian function. However, when c is larger than 7, then $\psi_0(x)$ and $\psi_1(x)$ have more than 99 per cent of their energy in $[-x_0, x_0]$, where the approximation [3.22] is satisfactory. In Fig. 1, we see the comparison between a Gaussian function (dotted line) with variance equal to $\sqrt{\frac{2}{c}}$ and $\psi_0(x)$ (solid line) with c equal to 7. $\psi_0(x)$ has been computed according to the approximation described in Frieden (1971).

3.3.2. Gaussian filtering and the heat equation

We consider here briefly an interesting analytic property of Gaussian filtering of images.

Gaussian filtering, i.e. the convolution of the image $I(x,y)$ (when $I(x,y)$ is bounded and continuous) with the Gaussian,

$$e^{\frac{-(x^2+y^2)}{2\sigma^2}}, \qquad [3.23]$$

can be seen as a solution at an appropriate time $t = \frac{\sigma^2}{2}$ of the two dimensional heat equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial u}{\partial t}, \qquad [3.24]$$

with the initial condition:

$$u(x,y,0) = I(x,y). \qquad [3.25]$$

This is because the "source solution" of the heat equation (Widder, 1975) is

$$k(x,y,t) = \frac{e^{\frac{-(x^2+y^2)}{4t}}}{\sqrt{4\pi t}}, \qquad [3.26]$$

with time playing the role of the variance, that is,

$$\sigma^2 = 2t. \qquad [3.27]$$

From Theorem 4.1 of Widder (1975) the solutions of the heat equation are *entire* functions of x and y. In other words the convolution of a continuous and bounded function with a Gaussian generates an entire function. This characterizes well the strong regularizing properties of the gaussian filter.

4. Differentiation stage

In this chapter, we will discuss the properties of some differential operators that have been proposed and used in edge detection. We first briefly consider directional derivatives in section 4.1. In section 4.2 we discuss properties of two second-order rotationally invariant differential operators: the Laplacian and the second derivative along the direction of the gradient. We stress here that it is unlikely that zero-crossings of one differential operator — such as the Laplacian — are sufficient for early vision.

The many two-dimensional differential operators that can be used for detecting sharp changes in intensity can be classified according to whether they are (a) linear or nonlinear, and (b) directional or rotationally symmetric. In this paper, we use the (somewhat inappropriate) terminology of zeros of a differential operator Df (f defined in $V \subset \Re^2$) in the sense of the locus of points of V such that $Df = 0$. This notion is different from the usual definition of the kernel of an operator D, that is, the set of function f such that $Df = 0$ in V.

4.1. Directional differential operators (DD)

The directional differential DD operators used in edge detection are the usual directional derivatives. The use of directional operators has been criticized (Hildreth, 1980) on the grounds that such operators lead to smearing of zero-crossing contours (see Fig. 11 of Hildreth 1980). In that case the vertical operator was implemented with an operator of $n \times m$ pixels. Smoothing was performed in both the orthogonal and the parallel direction to the filter's orientation. A correct implementation of a vertical derivative however consists of an operator of $1 \times m$ pixels. The smearing observed by Hildreth (1980), however, is not due to the use of a directional operator but to the distortion introduced by a too large width of the operator. The concomitant use of several directional derivatives has been proposed by several authors (Binford 1982; Canny 1983). Since in $\Re^2$ the directional derivative in any arbitrary direction can be expressed in terms of $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$, it is evident that in a noise-free image the use of more than two-directional derivatives is of no help at all. In a noisy image the use of several directional derivatives may be useful for increasing the signal-to-noise ratio.

We will see in a later paper that the use of just two narrow directional derivatives is sufficient to detect all edges detected by rotationally invariant differential operators or by a large set of directional derivatives.

4.2. Rotational invariant differential operators (RID)

Rotationally symmetric operators have several attractive features. Two of the most widely used operators of this class are the Laplacian (∇^2, which is linear) and the second directional derivative along the gradient ($\frac{\partial^2}{\partial n^2}$) which is nonlinear. We will derive in this section several properties of the two derivatives and especially of their zero-crossings. In particular, we derive a necessary and sufficient conditions on image intensities for the zero-crossings of the two derivatives to coincide.

4.2.1. Null space of the Laplacian and subharmonic functions

Certain classes of functions do not originate zero-crossings in the Laplacian: they are *harmonic* and *subharmonic* functions. Harmonic functions are the null space of the Laplacian operator. Interestingly, they are invariant with respect to heat diffusion and therefore do not change under convolution with a gaussian of any size (Yuille, pers. comm.). This property, however, is not stable. Another non trivial result is that any non-linear function ϕ of an harmonic function has zero-crossings at the locations of the inflection points of ϕ (Yuille, Poggio and Ullman, pers. comm.). Harmonic functions are non-generic, in the sense that a small perturbation destroys the harmonic property.

Subharmonic functions are such functions that the modulus of their Laplacian is everywhere positive (Daugman, 1984a). These functions are robust against small perturbations.

4.2.2. Cartesian and polar form

We just give the explicit representation of the two operators in cartesian and polar coordinates:

$$\nabla^2 f = f_{xx} + f_{yy} = \frac{\partial^2 f}{\partial \rho^2} + \frac{1}{\rho}\frac{\partial f}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2 f}{\partial \theta^2} \qquad [4.1]$$

$$\frac{\partial^2 f}{\partial n^2} = \frac{f_x^2 f_{xx} + 2 f_x f_y f_{xy} + f_y^2 f_{yy}}{f_x^2 + f_y^2} = [\frac{2}{\rho^2}\frac{\partial f}{\partial \rho}\frac{\partial f}{\partial \theta}\frac{\partial^2 f}{\partial \rho \partial \theta} + \frac{1}{\rho^4}\left(\frac{\partial f}{\partial \theta}\right)\frac{\partial^2 f}{\partial \theta^2} - \frac{1}{\rho^3}\frac{\partial f}{\partial \rho}\left(\frac{\partial f}{\partial \theta}\right)^2 + \left(\frac{\partial f}{\partial \rho}\right)^2 \frac{\partial^2}{\partial \rho^2}] \frac{1}{\left(\frac{\partial f}{\partial \rho}\right)^2 + \frac{1}{\rho^2}\left(\frac{\partial f}{\partial \theta}\right)^2}. \quad [4.2]$$

We also give the explicit representation for the second directional derivative in the direction orthogonal to the gradient:

$$\frac{\partial^2 f}{\partial n_\perp^2} = \frac{f_y^2 f_{xx} - 2 f_x f_y f_{xy} + f_x^2 f_{yy}}{f_x^2 + f_y^2} \quad [4.3]$$

Remark:

The representation in polar coordinates shows clearly that the two operators are rotationally symmetric, since their form does not change for a rotation of the coordinate system θ^*. We can now state

•*Characteristic Property of Rotationally Symmetric Operators.* A sufficient condition for an operator to be rotationally invariant is that θ appears only as derivative in the polar representation of the operator.

4.2.3. Simple properties of ∇^2 and $\frac{\partial^2}{\partial n^2}$

Marr and Hildreth (1980) had attempted to prove that in most cases the zero-crossings of the Laplacian coincide with the intensity edges. Since zeros of the second directional derivative along the intensity gradient are the natural definition of intensity edges, we are able to give here a more rigorous characterization of the problem, in terms of four simple properties.

(I) *If the image $f(x, y)$ can be represented as a function of only one variable, i.e., $f(x, y_o)$, the two operators ∇^2 and $\frac{\partial^2}{\partial n^2}$ are equivalent, i.e., $\frac{\partial^2}{\partial n^2} f = \nabla^2 f$.*

As a consequence, for $f(x, y_o)$ the zeros of $\frac{\partial^2 f}{\partial n^2}$ and of $\nabla^2 f$ coincide.

Property I is similar but *not* identical to the "linear variation" result of Marr and Hildreth (1980), which states that if f changes at most linearly along the edge direction ℓ, then $\nabla^2 f = \frac{\partial^2 f}{\partial l^2}$.

(II) *If $f_{yy} = f_{xy} = 0$ at P, when $\frac{\partial^2 f}{\partial n^2} = 0$, the zeros of $\frac{\partial^2 f}{\partial n^2}$ coincides with the zeros of $\nabla^2 f$.*

The assumptions on the image are here stronger than the condition of linear variation of Marr and Hildreth (1980), but are equivalent to the assumptions of their theorem 1: locally around the zero-crossing, f has the form $f(x, y) = ax + by + c$.

(III) *If $f(x, y) = f(\rho)$ is rotationally symmetric, $\nabla^2 f$ and $\frac{\partial^2 f}{\partial n^2}$ differ by the additive term $\frac{1}{\rho}\frac{\partial f}{\partial \rho}$.*

For circularly symmetric functions, the zeros of $\nabla^2 f$ are farther apart than the zeros of $\frac{\partial^2 f}{\partial n^2}$. This lack of localization by ∇^2 (for circularly symmetric patterns) can also be seen in the fact that zeros of ∇^2 (but not of $\frac{\partial^2}{\partial n^2}$) "swing wide" of corners.

(IV). *(a) $\frac{\partial^2}{\partial n^2}$ is nonlinear.*

(b) $\frac{\partial^2}{\partial n^2}$ neither commutes nor associates with the convolution, i.e.,

$$\frac{\partial^2}{\partial n^2}(g * f) \neq \left(\frac{\partial^2}{\partial n^2} g\right) * f \qquad [4.4]$$

$$\left(\frac{\partial^2}{\partial n^2} g\right) * f \neq g * \frac{\partial^2}{\partial n^2} f. \qquad [4.5]$$

(c) $\frac{\partial^2}{\partial n^2}$ *is a linear operator on* f, *if* $f = f(\rho)$, *but not shift invariant.*

(d)*The mean of* $\frac{\partial^2}{\partial n^2}$ *applied to a zero-mean function need not to be zero.*

4.2.4. Geometric characterization of the zeros of ∇^2 and $\frac{\partial^2}{\partial n^2}$

It is interesting to consider under which conditions the zeros of the Laplacian coincide with the zeros of the second directional derivative along the gradient. Zeros of the second directional derivative along the gradient are a natural way of characterizing and localizing intensity edges. Zeros of the Laplacian, however, are extensively used for their computational convenience. In this section we derive rigorous results that clarify completely this set of questions.

Let us consider the intensity surface represented as $X = (x, y, z)$, where $z = f(x, y)$ with $f \in C^r(D)$, $D \in \Re^2$ and $r > 2$.

The mean curvature of the surface X is

$$H = \frac{EN + GL - 2FM}{2g^2} = \frac{(1 + f_x^2)f_{yy} + (1 + f_y^2)f_{xx} - 2f_x f_y f_{xy}}{2g^3}, \qquad [4.6]$$

where

$$E = 1 + f_x^2 \quad F = f_x f_y \quad G = 1 + f_y^2, \qquad [4.7]$$

are the coefficients of the first fundamental form $I(dx, dy)$ (Lipschutz, 1969, Pogorelov, 1965), and

$$L = \frac{f_{xx}}{g} \quad M = \frac{f_{xy}}{g} \quad N = \frac{f_{yy}}{g}, \qquad [4.7]$$

with $g^2 = 1 + f_x^2 + f_y^2$, are the coefficients of the second fundamental form $II(dx, dy)$.

We use equations [4.2], [4.3] and [4.8] and the property

$$\nabla^2 f = \left(\frac{\partial^2}{\partial n^2} + \frac{\partial^2}{\partial n_\perp^2}\right) f \qquad [4.9]$$

for writing H in terms of ∇^2 and $\frac{\partial^2}{\partial n^2}$:

$$H = \frac{1}{2g^3}\left(g^2 \nabla^2 f - (\nabla f)^2 \frac{\partial^2 f}{\partial n^2}\right). \qquad [4.10]$$

We can now characterize the connection between the zeros of ∇^2 and the zeros of $\frac{\partial^2}{\partial n^2} f$:

Property V: If $\nabla f \neq 0$, the zeros of $\frac{\partial^2}{\partial n^2} f$ coincide with the zeros of ∇^2 *iff* the mean curvature H is zero.

Thus, only for surfaces with minimal curvature ($H = 0$), the zeros of $\frac{\partial^2 f}{\partial n^2}$ coincide with the zeros of $\nabla^2 f$ where the gradient of f is different from zero. Note that (M. Kass, personal communication) $\nabla^2 f$ has the same zeros as $\frac{\partial^2 f}{\partial n^2}$ where the curvature of the lines of level-crossings of the intensity image is zero. Recently, Berzins (1984) analyzed in detail the behavior of zeros of the Laplacian of a Gaussian filtered image around corner edges and edges with high curvature. He showed that the zeros of the Laplacian are displaced from the true edge by less than σ (the variance of the Gaussian filtering) when the radius of curvature is large compared to σ, and when the distance to the nearest sharp corner is large compared to $\frac{\Theta}{\sigma}$ (where Θ is the angle of the corner in radians). Note that eq. [4.10] shows that the difference between $\frac{\partial^2 f}{\partial n^2}$ and $\nabla^2 f$ is small if the mean curvature H is small. Smoothing the image with a two-dimensional filter reduces the curvature (and more so for larger-sized filters). Therefore, we may expect that in *filtered* images, $\nabla^2 f$ will perform almost as well as $\frac{\partial^2 f}{\partial n^2}$.

4.2.5. The normal curvature

The second directional derivative along the gradient has a simple interpretation in terms of the normal curvature along the gradient. The normal curvature K_n in the direction of the gradient is (Lipschutz, 1969)

$$K_n = \frac{L du^2 + 2M du dv + N dv^2}{E du^2 + 2F du dv + G dv^2} \qquad [4.11]$$

with du and dv as direction numbers. Setting $du^2 + dv^2 = 1$, the direction numbers along the gradient are

$$du = \frac{f_x}{|\nabla f|} \qquad [4.12]$$

$$dv = \frac{f_y}{|\nabla f|}. \qquad [4.13]$$

Thus, equations [4.11] and [4.13], together with equations [4.6] and [4.7], lead to

$$K_n = \frac{1}{g^3} \frac{\partial^2}{\partial n^2} f. \qquad [4.14]$$

In particular, it follows

Property VI

The second directional derivative along the gradient and the normal curvature in the direction of the gradient have the same zeros when $|\nabla f| \neq 0$.

Our geometrical characterization of the gradient and the second derivative along the gradient is completed by Appendix 3, that gives the geodesic curvature of the curve directed along the gradient. For surfaces of revolution the geodesic curvature of such lines is always zero.

The operator $\frac{\partial^2}{\partial n^2}$ and the normal curvature in the direction of the gradient K_n are not defined when $|\nabla f| = 0$. In this case, the direction of the gradient is underdetermined, although the Hessian can of course be diagonalized (determining the principal directions). Thus, $\frac{\partial^2}{\partial n^2}$ has the disadvantage with respect to ∇^2 that it is not defined everywhere.

4.2.6. Potential biological consequences

A natural question arising from these comparisons is: which derivative operators are used by the human visual system? It is obvious from the earlier sections that several different derivatives possibly at different scales have to be used for efficient edge detection. It would be very strange if the human visual system would make use of only one differential operator. The important questions is therefore which operators or combinations thereof are used in different visual tasks and under different conditions. Zero-crossings in the output of directional second derivatives approximated by the difference of one-dimensional Gaussians (DOG) were suggested by Marr & Poggio (1977) in their theory of stereo matching. Marr & Hildreth (1980) later proposed the rotationally symmetric Laplacian $\nabla^2 G$ (approximated by a rotationally symmetric DOG) for edge detection and for stereo matching. Psychophysical evidence does not rule out either of these schemes. Physiology shows that a class of retinal ganglion cells performs a roughly linear operation quite similar to the convolution of the image with the Laplacian of a Gaussian. Data on cortical cells are still somewhat contradictory on whether some simple cells may perform the equivalent of a linear directional derivative operation, or instead, signal the presence of a zero-crossing of the rotationally symmetric $\nabla^2 G$.

On physiological grounds, it seems unlikely that retinal cells could perform the rotationally symmetric nonlinear $\frac{\partial^2}{\partial n^2}$ operation, although not all classes of ganglion cells have been tested properly to allow a firm conclusion. In particular, one-dimensional and rotationally symmetric patterns are customarily used as stimuli for physiological experiments. In the first case $\frac{\partial^2}{\partial n^2}$ and ∇^2 are equivalent, whereas in the second case, they may be distinguishable only quantitatively. Let us now consider three classes of psychophysical experiments.

(I) An interesting possibility for distinguishing the Laplacian from the directional second derivative on the basis of physiological or psychophysical experiments is suggested by the observation that the zero-crossings of the Laplacian "swing wide" of gray-level corners. In particular, the zero-crossings associated with an elongated black bar, for example, coincide for ∇^2 and $\frac{\partial^2}{\partial n^2}$, whereas they differ in the case of a circular black disk. Hyperacuity experiments may allow one to distinguish the two cases. Notice that both operators are linear in this case. They associate therefore with Gaussian convolution ($G = e^{\frac{-x^2}{2\sigma^2}}$). The corresponding point-spread functions are

(a) for the one-dimensional, $f(x)$:

$$\frac{\partial^2}{\partial x^2} G = \frac{1}{\sigma^2}\left(\frac{x^2}{\sigma^2} - 1\right) e^{\frac{-x^2}{2\sigma^2}} \qquad [4.15]$$

(II) for the two-dimensional $f(\rho)$

$$\nabla^2 G = \frac{2}{\sigma^2}\left(\frac{\rho^2}{2\sigma^2} - 1\right) e^{\frac{-\rho^2}{2\sigma^2}} \qquad [4.16]$$

$$\frac{\partial^2 G}{\partial n^2} = \frac{1}{\sigma^2}\left(\frac{\rho^2}{\sigma^2} - 1\right) e^{\frac{-\rho^2}{\sigma^2}} \qquad [4.17]$$

where σ is the standard deviation of the Gaussian function. Let us call w the diameter of the central region of these masks, i.e., the distance between the central zeros. w_{1D} denotes the diameter for the one-dimensional case and w_{2D} for the two-dimensional case. It is easy to see that the second directional derivative has $w^d_{1D} = w^d_{2D}$ whereas this is not true for the Laplacian $w^\ell_{1D} \neq w^\ell_{2D}$. From (a) and (b) we get

$$\begin{aligned} w^d_{2D} = w^d_{1D} = w^\ell_{1D} &= 2\sigma \\ w^\ell_{2D} &= 2\sqrt{2}\sigma \end{aligned} \qquad [4.18]$$

A possible psychophysical test is:

•If zero-crossings in the Laplacian are used by our visual system to estimate position of edges, the apparent width of a narrow 1-D bar and of a small circle (with equal physical widths) should be different—the bar should appear smaller. This is not expected if the second directional derivative is used.

(II) There are classes of intensity edges that generate zeros in $\frac{\partial^2}{\partial n^2}$ but not in ∇^2. An example is given by:

$$I(x,y) = (1 + e^{\beta y})\frac{e^{\alpha x}}{1 + e^{\alpha x}} \qquad [4.19]$$

which, with appropriate values of β does not satisy $\nabla^2 I = 0$ for any $y \geq 0$. It is possible, however, to find solutions to $\frac{\partial^2}{\partial n^2} I = 0$. Thus, the edge I could again be used to discriminate psychophysically between ∇^2 and $\frac{\partial^2}{\partial n^2}$.

More in general, functions $h \in C^2$ in a certain region D such that $\nabla^2 h \geq 0$ in D are called subharmonic, as we discussed earlier. These functions do not have zero-crossings of the Laplacian (Daugman, 1984a), but generally zero-rossings of $\frac{\partial^2}{\partial n^2}$ are present. There are special cases, however, in which both $\frac{\partial^2}{\partial n^2}$ and ∇^2 do not have zeros. An example is given by $f(x) = \cos x + bx^2$ with $\nabla^2 f = \frac{\partial^2}{\partial n^2} f = -\cos^2 x + 2b$, which does not have any zero-crossings if $b > \frac{1}{2}$. It would be interesting to test this kind of pattern both psychophysically and physiologically (controlling carefully for nonlinearities in the phototransduction).

III) As we mentioned earlier in this chapter, harmonic functions cannot be characterized in terms of the zero-crossings of their Laplacian. Worse yet, any image is characterized uniquely by zero-crossings of the Laplacian (across gaussian scales, see chapter 6) modulus any harmonic function. Psychophysical experiments that measure the detectability of edges in subharmonic patterns are difficult to interpret, because they would give a clear answer only if the Laplacian were the only differential operator in the human visual system, a very unlikely possibility. Furthermore, harmonic functions are unstable against small perturbations, making difficult to control for non-linearities in the display and in the transduction process.

5. Geometrical properties of edge contours

In this section, we will discuss geometrical properties of edge contours obtained by different methods. We will show that edges derived through rotational operators are generally smooth, closed curves, while edges obtained with directional operators do not have such special geometrical properties.

In many edge detection schemes, as we discussed in the Introduction, the image $I(x,y)$ is first filtered and then a second order differential operator D^2 is applied to the filtered image $\hat{I}(x,y)$. Edges are identified in correspondence to the zero-crossings of $D^2\hat{I}(x,y)$. In other cases, edges are identified as extrema of some derivative of the filtered image. Again they may correspond to zero-crossings of a higher order derivative. In this way, the first part of edge detection provides a *compact* and possibly *complete* representation of intensity changes (see chapter 6).

Therefore, it is important to analyze theoretically geometrical properties of the locus of points defined by

$$D^2\hat{I}(x,y) = 0, \qquad [5.1]$$

where $\hat{I}(x,y)$ is the filtered image and D^2 can be a RID or a DD operator. We first recall in the next two sections the notions of transversality (Abraham and Robbin, 1967, Poston and Stewart, 1976) and of Morse functions (Poston and Stewart, 1976). In section 5.4 we will classify the types of zero-crossings contours that can appear in images.

5.1. Transversality and zero-crossing (z.c.)

A curve (or a surface) S_1 meets a curve (or a surface) S_2 in P *transversally* when the tangent space TS_1 to S_1 in P and the tangent space TS_2 to S_2 in P have locally around P an empty intersection. More generally, two subspaces U, V of $\Re^n$ are transverse if they meet in a subspace whose dimension is as small as possible. From this definition, it follows that the surface $S_f = (x, y, f(x,y))$ meets the surface $S_o = (x, y, 0)$ in $\hat{P} = (\hat{x}, \hat{y}, 0)$ transversally *iff* in $(\hat{x}, \hat{y})$

$$|\text{grad} f| \neq 0. \qquad [5.2]$$

The isotopy theorem (Thom, 1954) shows that transversal intersections are structurally stable. The converse is also true in that non-transversal intersections are structurally unstable. Transversality (and the implicit function theorem) indicates that if S_f meets S_o transversally in $\hat{P}$ then the intersection of S_f and S_o around $\hat{P}$ is a smooth curve.

The previous result is only local. Globally we find that if $f(x,y)$ is defined in the compact domain $V \in R^2$ whose boundary is δV and if S_f always meets S_o transversally, then the intersection of S_f and S_o consists of:

(a) smooth closed curves $\Gamma_c \in V$.

(b) smooth curves Γ_t that terminate in δV.

In other words, transversality of zero-crossings means that *zero-crossing contours are closed curves or curves that terminate at the boundary of the image.*

5.2. Closed and open contours of z.c.

From the previous section a necessary and sufficient condition for transversality in P is:

$$|\text{grad} D^2 \hat{I}(x,y)|_P \neq 0. \qquad [5.3]$$

A preliminary condition required by eq. [5.3] is that $D^2\hat{I}(x,y)$ is a differentiable function. This condition is obviously met if $\hat{I}(x,y)$ is analytic (or entire or band-limited). But we have already stressed that it is safer to suppose that the original image $I(x,y)$ is a piece-wise continuous function or belongs to C^n, with n not known *a priori*. If we filter the original image $I(x,y)$ with an appropriate rotational filter, then $\hat{I}(x,y)$ is analytic both in x and y, and $D^2\hat{I}(x,y) = 0$ defines a differential function. On the other hand, if we use a directional filter f, for example along x, we have

$$\hat{I}(x,y) = I(x,y) * f(x) \qquad (5.4)$$

and there is no reason for $\hat{I}(x,y)$ to be a three times differentiable function of y. Therefore, if the original image has been filtered with a directional operator only, it is possible that the zero-crossings of $D^2\hat{I}(x,y)$ may not be smooth curves.

5.3. Morse functions

A function $f: R^2 \mapsto R$ is called a Morse function if at its critical point (i.e., points where $\text{grad} f = 0$), the Hessian is nondegenerate. Morse functions have the following properties:

(a) Suppose that $f(\hat{x},\hat{y}) = 0$ and $|\text{grad}f|_{\hat{P}} = 0$ with $\hat{P} = (\hat{x},\hat{y})$, but the Hess $(f)_{\hat{P}}$ is non degenerate. Thus, there is a smooth local change of coordinates around $\hat{P}|\text{grad}(x,y)_{\hat{P}}|$, such that f takes the exact form

$$f(x,y) = \frac{1}{2}\frac{\partial^2 f}{\partial x^2}|_{\hat{P}}x^2 + \frac{\partial^2 f}{\partial x \partial y}|_{\hat{P}}xy + \frac{1}{2}\frac{\partial^2 f}{\partial y}|_{\hat{P}}y^2. \qquad [5.5]$$

(b) A small enough perturbation of a Morse function f can always be expressed in the same form as the original f by a change of coordinates and of scale.

Property (a) says that around $\hat{P}$ the function f has the behavior of the quadratic form induced by the Hessian. Property (b) is a kind of structural stability property. A basic property of Morse functions is that they are dense so that, if f is a non-Morse function, then an arbitrary small perturbation of f makes f a Morse function (obviously the perturbation must not vanish at the critical points). This is the reason of the importance of Morse functions here: we can always assume that images are Morse functions (especially because of the unavoidable noise).

5.4. Classification of z.c.

We now analyze the geometrical properties of the z.c. contours, i.e., the locus of points defined by

$$D^2\hat{I}(x,y) = h(x,y) = 0. \qquad [5.7]$$

(a) If $h(x,y)$ is not a smooth function of x and y (at least C^1), the implicit function theorem cannot be used and the z.c. may be isolated points, i.e., segments of intersecting curves and 2-D regions.

(b) If $h(x,y)$ is a smooth function of x and y and if in $\hat{P} = (\hat{x},\hat{y})$ we have

$$h(\hat{x},\hat{y}) = 0 \quad \text{and} \quad |\text{grad}h(x,y)|_{\hat{P}} \neq 0,$$

then $h(x,y)$ has in $\hat{P}$ a "transversal zero-crossing", which is a smooth curve.

(c) If $h(x,y)$ is a smooth function and in $\hat{P}$ we have

$$h(\hat{x},\hat{y}) = 0 \quad \text{and} \quad |\text{grad}h(x,y)_{\hat{P}}| = 0 \qquad [5.8]$$

but around $\hat{P}$, $h(x,y)$ we find

$$h(x,y) = ax^2 + bxy + cy^2 + O(x^n y^m) \qquad n+m=3, \qquad [5.9]$$

where $a = \frac{1}{2}f_{xx}|_{\hat{P}}$, $b = f_{xy}|_{\hat{P}}$, $c = \frac{1}{2}f_{yy}|_{\hat{P}}$. The zero crossing $\hat{P}$ is:

(i) an elliptic z.c., if Hess $h(x,y)|_{\hat{P}} > 0$ (see Fig. 2A).

(ii) a hyperbolic z.c., (saddle point) if Hess $h(x,y)|_{\hat{P}} < 0$ (see Fig. 2B).

(iii) a parabolic z.c., if Hess $h(x,y)|_{\hat{P}} = 0$ but a, b and c are not identical to zero (see Fig. 2C).

(d) If $h(x,y)$ is a smooth function and if in $\hat{P}$ we have

$$h(\hat{x},\hat{y}) = 0 \quad |\text{grad}f(x,y)|_{\hat{P}} = 0,$$

and in $\hat{P}$, $h(x,y)$ depends on the third order terms,

$$h(x,y) = \alpha x^3 + \beta x^2 y + g x y^2 + \delta y^3 + O(x^n y^m), \quad n+m=4, \tag{5.10}$$

where the coefficients α, β, j and δ are obtained by the Taylor expansion. It is easy to see that the set of points

$$R_A = \{(x,y): \alpha x^3 + \beta x^2 y + \gamma x y^2 + \delta y^3 = 0\} \tag{5.11}$$

are straight lines. The z.c. lines may be:

(i) an elliptic umbilic, if R_A consists of three lines (see Fig. 2D).

(ii) a hyperbolic umbilic, if R_A consists of a single real line (see Fig. 2E)

(iii) a parabolic umbilic, if R_A consists of three lines, two of which are coicident (see Fig. 2F)

(iv) a symbolic umbilic, if R_A consists of three coincident lines.

(e) If $h(x,y)$ is a smooth function and in $\hat{P}$ we have

$$h(\hat{x},\hat{y}) = 0$$

and in $\hat{P}$, $h(x,y)$ depends on the fourth order terms, the z.c. lines have a complex shape that can be analyzed using results of Poston & Stewart (1976).

Bifurcations of zero-crossings

The isotopy theorem (Thom, 1954, Abraham and Robbin, 1967) shows that transversal intersections are structurally stable, i. e. that "transversal zero-crossings" are structurally stable: their topological properties do not change if the size of the filter is slightly changed.

If $f(x,y)$ is a Morse function then S_f may meet S_o non-transverally, and these intersections are not structurally stable (observe that Morse functions are structurally stable but not their intersections with S_o). If f is a Morse function, then S_f may meet S_o non-transversally at elliptic points and hyperbolic points. These intersections are not structurally stable and may change their topological properties for small perturbations of f. More specifically we may have two bifurcations:

(i) Elliptic z.c. At elliptic z.c., a small perturbation of f may lead to the disappearance of the z.c. or to the appearance of a contour of z.c. constituted by a closed curve.

(ii) hyperbolic z.c. At hyperbolic z.c., which consists of the intersection of two curves any small perturbation leads to the breaking of the intersection of the two curves and the appearance of two disjoint curves.

These are the two bifurcations that may appear when $h(x,y)$ is a Morse function. Interestingly enough, the zero-crossing contours obtained with real images (which will be explored in a later paper) can be classified as type (b) and (c) of the previous section; Morse functions can have z.c. only of type (b) and (c). The two types of bifurcation, that may originate with Morse functions are illustrated in Fig. 3A and 3B, respectively (see also Koenderink and van Doorn, 1979). Yuille and Poggio (1983a, 1983b) have shown that (if Gaussian filtering is used) when the scale of the filter is changed (i.e. σ), the second type of bifurcation may appear either when σ is increased or decreased, but the first type of bifurcation only occurs when σ is increased. Thus, the Gaussian filter forbids creation of a zero-crossing contour from an elliptic z.c. for increasing σ. It is important to note that all these topological properties are also valid for level-crossings. Thus setting a threshold in the output of the

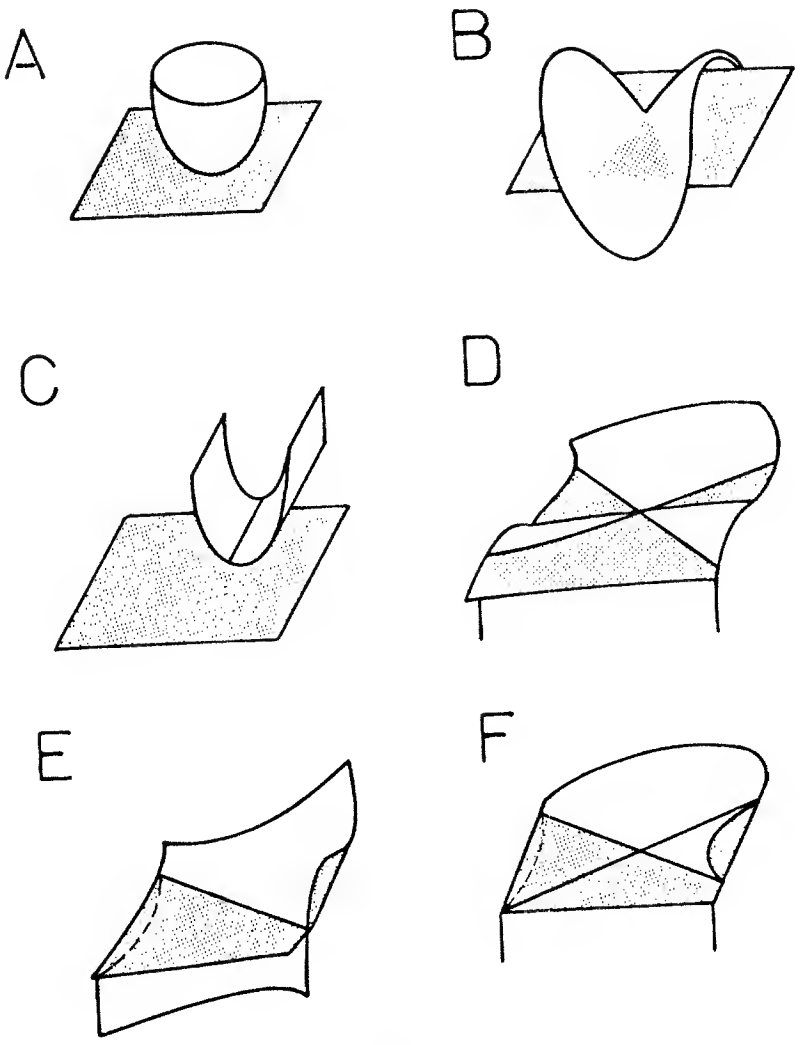

Figure 3 *The zero-crossing points may be of the elliptic (A), hyperbolic (B), parabolic (C) type; the zero-crossing lines can also be an elliptic umbilic (D), a hyperbolic umbilic (E) or a parabolic umbilic (F). See text.*

filtering and derivative operation preserves all topological and geometrical properties of zero-crossings.

In summary, we have characterized the geometrical properties of zero-crossing contours: these properties — for instance the fact that zero-crossing contours are closed — may be exploited in various ways in edge detection and even in stereo- or motion-matching.

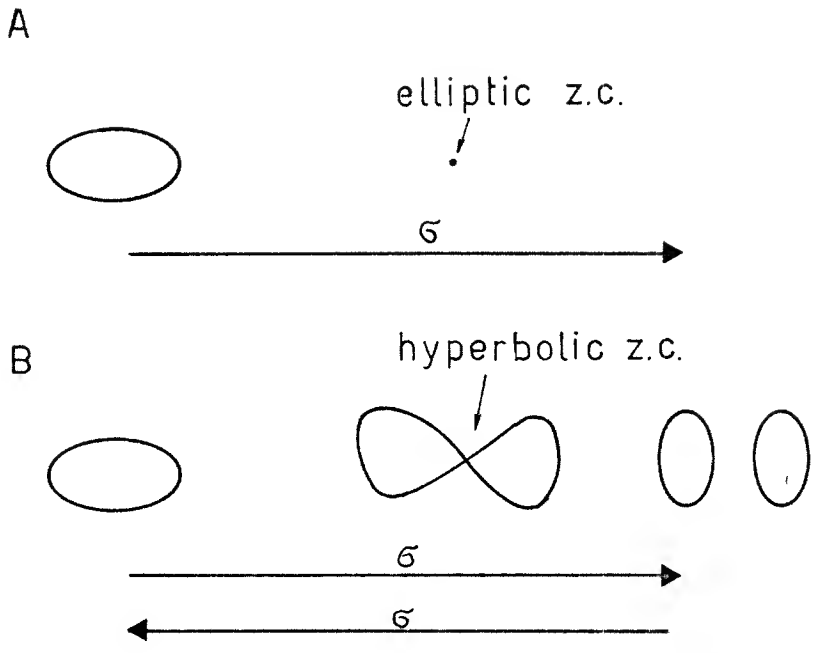

Figure 4 *The two types of bifurcations that can occur for increasing (left to right) and decreasing (right to left) σ in the case of Morse functions. See text.*

6. EDGE CONTOURS AND FILTER SCALE

As we have seen, differential operations on sampled images require the image to first be smoothed by filtering. The filtering operation introduces an arbitrary parameter – *the scale* of the filter, e.g., the standard deviation for the Gaussian filter. In computer vision, the advantages of using several scales of filtering was realized quite early on, and this was supported by evidence suggesting the presence of filters of several sizes in the human visual system (Rosenfeld, 1982; Marr, 1976; Marr and Poggio, 1979; Marr and Hildreth, 1980). More recently, Witkin (1983; see also Stansfield, 1980) introduced a scale-space description of zero-crossings which gives the position of the zero-crossing across a continuum of scales, i.e., sizes of the Gaussian filter (parametrized by the σ of the Gaussian). The signal—or the result of applying to the signal a linear (differential) operator—is convolved with a Gaussian filter over a continuum of sizes of the filter. Zero- or level- crossings of the filtered signal are contours on the $x - \sigma$ plane and surfaces in the x, y, σ space. Witkin proposed that this concise map can be effectively used to obtain a rich and qualitative description of the signal. Yuille and Poggio (1983a, 1983b) — who called the maps of zero-crossings across scales *fingerprints* — have established interesting relationships between multiresolution analysis, the Gaussian filter and zero-crossings of filtered signals. Their main results are

two theorems:

(a) zero- and level-crossings of an image filtered through a Gaussian filter have *nice* scaling properties, i.e., a simple behavior of zero-crossings across scales. Zero-crossings are not created as the scale increases. The Gaussian filter is the only filter that has this *nice* scaling behavior (see also Babaud, Witkin and Duda, 1983).

(b) The map of the zero-crossings across scales determines the filtered signal uniquely for almost all signals in the absence of noise. The scale map obtained by Gaussian filters is thus a *complete representation* of the image. This result applies to level-crossings of any arbitrary linear differential operator of the Gaussian (modulus the null space of the differential operator and provided there are at least two zero-crossing contours), since it applies to functions that obey the diffusion equation.

The first result sheds some light on the properties of zero-crossings and level-crossings at different scales with the Gaussian filter. It supports the use of the Gaussian filter in a multiresolution edge detection scheme. Reconstruction of the signal is, of course, not the goal of early signal processing. Symbolic primitives must be extracted from the signals and used for later processing. The second result implies that scale-space fingerprints are complete primitives, that capture the whole information in the signal and characterize it uniquely. Subsequent processes can therefore work on this more compact representation instead of the original signal (see Asada and Brady, 1984).

The second theorem has theoretical interest in that it answers the question of what information is conveyed by the *edges* identified with zero- and level-crossings of multiscale Gaussian filtered signals. It is furthermore interesting that this *complete* representation happens to coincide with the basic scheme for edge detection discussed in this paper. From this point of view it can be argued that the fingerprint representation makes explicit exactly the information that is needed on physical grounds, i.e., it makes explicit edges in the image.

It may be asked at this point what the right sequence is for the two steps of differentiation and filtering. For linear operators the order is of course immaterial, since they commute. It is not so for nonlinear operators, such as the directional derivative along the gradient. The regularization argument for the filtering step implies that filtering at one scale must precede the differentiation operation. The computation of different scales requires filtering at a range of resolutions *after* differentiation. The reason is that the theorems of Yuille and Poggio (1983a, 1983b) hold true even for the identity operator, but are not necessarily valid if filtering is performed before a *nonlinear* differential operation. In particular, Gaussian scaling after the nonlinear directional derivative along the gradient does not have a *nice* scaling behavior. Thus filtering as a regularizing operator must be performed first at one scale and filtering at different scales must be performed after the differential operation. For linear differential operators, this is equivalent to a multiscale filtering either before, after, or together with the differential operation (e.g. the Laplacian of the Gaussian).

7. OVERVIEW OF SOME EDGE DETECTORS

In this section, we will briefly compare our main conclusions with several edge detectors presented in the literature. Our review is neither intended to be exhaustive nor does it aim to present edge detectors in full detail.

7.1. Difference of boxes (DOB)

Binford and coworkers (Herskovitz and Binford, 1970; Horn, 1972; Binford, 1981) have suggested the use of support-limited filters in the filtering step of edge detection. They have used the Haar function [3.8b] in directional filtering or a difference of functions of the type [3.8a] for rotational filtering. There are two problems using this approach:

(1) Filtering with support-limited functions does not regularize the image intensity profile; therefore the use of any differential operator is unsafe.

(2) A strictly support-limited filter, such as a DOB, cannot be correctly sampled, and it is very difficult to obtain a good digital representation.

7.2. Shanmugam, Dickey and Green

Shanmugam, Dickey and Green (1979) looked for a linear, band-limited operator that would yield maximal output energy within a given spatial interval in the vicinity of the edge. No explicit reference was made to a differentiation step in edge detection. They proposed that the optimal filter for an ideal edge $S(x)$, has a Fourier transform

$$F_{op}(\omega) = \begin{cases} k_1 \omega \psi_1\left(\frac{\omega x_a}{\Omega}, c\right) & |\Omega| \leq r \\ 0 & |\Omega| > r \end{cases}, \qquad [7.1]$$

where k_1 is a constant, $\psi_1(x, c)$ is a linear prolate function (see section 3.1). This edge detector performs very poorly on localization and has the intrinsic feature of giving two maxima of energy in the output of the response to an edge. The reason is simply that using an even filter such as eq. [7.1] which has the same shape of $\omega^2 \psi_o(\frac{\omega x_a}{\Omega}, c)$, edges are located at the zero-crossing of the output and not at the extrema. Moreover, these authors use properties of linear prolate functions (their eq. 1) to derive their optimal filter which are valid in 1-D, but not in 2-D when linear prolate functions are extended in 2-D by rotation. In addition, their asymptotic approximation to the optimal filter was incorrect, as shown by Lunscher (1983).

7.3. Marr and Hildreth

Marr & Hildreth (1980), and Hildreth (1980), extending the work of Marr and Poggio (1979), have proposed an edge detection scheme based on a filtering step consisting of a 2-D symmetric Gaussian followed by the localization of zero-crossings of $\nabla^2 \hat{I}(x, y)$, where $\hat{I}(x, y)$ is the filtered image. This edge detector performs rather well, but its optimality was not rigorously proved. Indeed,

(1) $\frac{\partial^2}{\partial n^2}$ in many instances achieves a better localization than ∇^2, particularly for rounded edges with large curvature.

(2) The use of directional filters and directional derivatives when performed correctly does not give rise to the problems that forced Marr and Hildreth to reject such edge detection schemes (see section 4.1). The use of two directional filters with directional derivatives may be as efficient as the Marr-Hildreth scheme, with the advantage of not introducing spurious edges that appear with rotational filtering because of the closure property of z.c. contours (see section 5).

7.4. Haralick

Haralick (1980, 1981, 1982) has proposed a scheme for edge detection in which a pixel is marked as a step edge pixel if, in its neighborhood, there is a zero-crossing of the second directional derivative taken in the direction of the gradient. Haralick, in order to evaluate the derivatives he approximates, interpolates the sampled intensity values with discrete Chebychev polynomials. There is no explicit mention of a filtering step. Canny (1983), however, has shown that the above procedure is practically equivalent to using a filtering step (in our terms, a regularization step) before differentiation. The type of equivalent filter depends on the set of approximating functions and on the degree of differentiation required.

7.5. Canny

Canny (1983) has investigated the desirable properties of an optimal edge detector, based on efficiency of detection and reliability in localization. We have already seen that detection

of an ideal step edge is favored by broad filters while localization is favored by small filters. Canny has shown through variational methods that the optimal odd filter $f_{op}(x)$ (according to his criteria) in the 1-D case is the linear combination of four exponentials.

Interestingly $f_{op}(x)$ is very close to $xe^{\frac{-x^2}{\sigma^2}}$, which is the optimal odd filter from the point of view of minimal uncertainty. The treatment of Canny may also be seen as a well-founded justification for the use of filters with minimal uncertainty, because simply by first changing some constraints in his variational approach it is possible to obtain the second Hermite function.[2] Canny's procedure for finding two-dimensional step edges and other types of edges uses directional operators of varying width, length and orientation. This procedure, which includes as an essential part an appropriate thresholding, works remarkably well on real images. His justification of the choice of directional operators is not completely satisfactory. Indeed:

(1) For 2-D images, Canny uses two alternative differential operators, either $\frac{\partial^2}{\partial n^2}$ (see section 4.2 and Havens and Strickwerda, personal communication) or directional operators. The preference for directional operators originates from his one-dimensional treatment of the problem. The optimal filter is chosen to be an antisymmetric function, because it is designed to detect maxima. Therefore the corresponding 2-D operator is not rotationally invariant, suggesting the use of directional operators for 2-D images. The output of directional operators can be directly used in the adaptive threshold scheme used by Canny, offering advantages with respect to the symmetric operator $\frac{\partial^2}{\partial n^2}$.

(2) As already mentioned in section 4.1, to obtain all edges in a 2-D image it is sufficient to use only two different directional derivatives. The use of more than two orientations is useful only to increase the signal-to-noise ratio, but is not required for edge detection in a noise-free 2-D image.

8. DISCUSSION

We will now summarize the main points of our analysis of edge detection.

A. The first step in edge detection, after sampling of the image, consists of a filtering stage followed by a differentiation stage. Filtering has the main function of regularizing the ill-posed nature of edge-detection and should be performed before the differentiation operation. Filtering for the purpose of multiresolution analysis should be performed after the differentiation operation, when nonlinear differential operators are used.

B. To be physically realizable, digital filters should be represented with a good approximation by a finite sequence of samples of points. From this point of view, a Gaussian or the first linear prolate ($\phi_o(x,c)$) function are practically equivalent. Filtering with prolate functions regularize "more" the image (the image becomes entire *and* band-limited), whereas using a Gaussian the image becomes only entire. The Gaussian filtering, however, has two advantages over prolate functions:

(i) It does not create z.c. when the size of the filter is increased (see section 6).

(ii) In 2-D, the Gaussian decomposes into the product of 1-D Gaussians: as a consequence, it is particularly easy to reduce drastically the amount of computations involved in its use.

C. Filtering of the image with a rotationally symmetric filter insures with high probability the closure of z.c. contours (see section 5). Filtering the image with directional filters does not ensure closed z.c. contours. Localization, however, is more accurate.

D. Several types of derivatives at different scales may be needed for detecting and labeling intensity changes under the most general conditions. In the differentiation step, directional

[2]Recently Spacek and Brady have investigated split-gaussian filters similar to Canny's but with poorer signal-to-noise ratio and better localization.

derivatives in only two directions are necessary, when DD operators are used. When RID operators are used, $\frac{\partial^2}{\partial n^2}$ performs better then ∇^2 in localization, but $\frac{\partial^2}{\partial n^2}$ has the disadvantage of not commuting with the convolution.

E. In order to characterize the types of intensity changes in the image in terms of the physical properties that have originated them, it is useful to have a set of hierarchical symbolic descriptions. The lowest symbolic description uses as a substrate the associated fingerprints of the image, containing the map of zero crossings and their slope at different scales, and provides a local labeling of edges still in terms of image data. The final symbolic description must label edges in terms of the properties of the physical surfaces that originate the intensity changes, and therefore as object boundaries, shadows, reflexes, changes in texture, specular reflections, etc. This final representation of the type of the primal sketch is obtained using high level knowledge and geometrical reasoning from lower symbolic descriptions.

In later papers, we will evaluate performance of different filters and different operators in real images, and we will outline a theory of a symbolic description of edges.

Acknowledgement: We are grateful to A. Yuille, A. Verri, M. Kass for useful discussions and suggestions. M. Bertero first pointed out to us that differentiation is an ill-posed problem. M. Brady, J. Canny, E. Grimson, M. Kass, H. Voorhees, W. Richards and especially E. Hildreth read the manuscript and provided tremendously useful and poorly implemented suggestions. Carol Bonomo typed the math, edited the English, and generally managed to look quite busy, even if she wasn't.

9. References

Abraham, R. and Robbin, J. *Transversal mappings and flows.* W. A. Benjamin, Inc., New York, 1967.

Achieser, N. I. *Theory of Approximation.* Frederick Ungar Publishing Co., New York, 1956.

Ahlberg, J. H., Nilson, E. H. and Walsh, J. L. *The Theory of Splines and their Applications,* (Vol. 38 in: MATHEMATICS IN SCIENCE AND ENGINEERING), Academic Press, New York, 1967.

Asada H. and Brady, M. The curvature primal sketch. A.I. Memo 758, MIT, 1984.

Ballard, D. H. and Brown, C. *Computer Vision*, Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

Babaud, J., Witkin, A. and Duda, R., "Uniqueness of the Gaussian kernel for scale-space filtering," Fairchild TR 645, Flair 22, 1983.

Berzins, V. " Accuracy of Laplacian edge detectors." Computer Graphics and Image Processing, 27, 195-210, 1984.

Binford, T. O. " Inferring surfaces from images." *Art. Int.*, 17, 205-244, 1981.

Binford, T. O. " Survey of model-based image analysis systems." *Int. J. Robotics Res.*, 1, no. 1, 18-64, 1982.

Boas, R. P. *Entire functions.* Academic Press, New York, 1954.

Borsellino, A., Poggio, T. "Correlation and Convolution algebras", *Kybernetik*, 13, 113-122, 1973.

Brady, J. M. " Computational Approaches to Image Understanding" *Computing Surveys*, 14, 3-71, 1982.

Canny, J. F. *Finding edges and lines.* MIT Technical Report No. 720, 1983.

Davis, L. "A survey of edge detection techniques." *Computer Graphics and Image Processing*, **4**, 248-270, 1975a.

Davis, P. J. *Interpolation and Approximation.* Dover Publications, Inc., New York, 1975b.

Daugman, J. G. "Six formal properties of two-dimensional anistropic visual filters: Structural principles and frequency/orientation selectivity." *IEEE Systems, Man and Cybernetics*, in press, 1984.

Daugman, J. G. "Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters." *J. Opt. Soc. Am*, in press, 1984a.

Frieden, B. R. "Linear and circular prolate functions", in **Progress in Optics IX**, ed. E. Wolf, North-Holland Publishing Co., Amsterdam, 312-408, 1971.

Greville, T. N. E. "Introduction to spline functions", in **Theory and Applications of Spline Functions**, ed. T. N. E. Greville, Academic Press, New York – London, 1-36, 1969.

Haralick, R. M. "Edge and region analysis for digital image data." *Compt. Graphics & Image Proc*, 12, 60-73, 1980.

Haralick, R. M. "The digital edge." *Proc. 1981 Conf. on Pattern Recognition & Image Processing*, Dallas, Texas, 285-294, 1981.

Haralick, R.M. "Zero-crossings of second directional derivative edge operator." *SPIE Proc. on Robot Vision*, Arlington Virginia, 1982.

Hermuth, H. H. *Transmission of information by orthogonal functions.* Springer-Verlag, Berlin, 1972.

Herskovitz, A. and Binford, T. O. "On boundary detection." *A. I. Memo 183*, MIT, 1980.

Hildreth, E. C. "Implementation of a theory of edge detection." *A. I. Memo 579*, MIT, 1980.

Horn, B. K. P. "The Binford-Horn edge finder." *A. I. Memo 285*, MIT, 1972.

Koenderink, J.J. and van Doorn, A. J. "The structure of two-dimensional scalar fields with applications to vision. " Biol. Cyb., 1979.

Kolmogorov, A. and Fomine, S. *Elements de la theorie des functions et de l'analyse fonctionelle.* Editions MIR, Moscov, 1974.

Landau, H. J. and Pollack, H. O. "Prolate spherical wave functions, Fourier analysis and uncertainty–II." *Bell Syst. Tech. J.*, **40**, 65-84, 1961.

Lipschutz, M. M. *Differential geometry.* McGraw-Hill, New York, 1969.

Lowe, D. G. and Binford, T. O. "Interpretation of geometric structure from image boundaries." *SPIE*, **281**, 224-231, 1981.

Lunscher, W.H.H. "The asymptotic Optimal Frequency Domain Filter for Edge Detection." *IEEE Trans. PAMI*, **6**, 678-680, 1983.

Marr, D. C. and Hildreth, E. C. "Theory of edge detection." *Proc. R. Soc. Lond. B*, **207**, 187-217, 1980.

Marr, D., Poggio, T. "A computational theory of human stereo vision." *Proc. R. Soc. Lond. B*, **204**, 301-328, 1979.

Oppenheim, A. V. and Johnson, D. H. "Discrete representation of signals." *Proc. IEEE*, **60**, no. 6, 681-691, 1972.

Papoulis, A. *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 1965.

Papoulis, A. *The Fourier Integral and its Applications*, McGraw-Hill, New York, 1962.

Pavlidis, T., *Structured Pattern Recognition*, Springer-Verlag, New York, 1977.

Poggio, T., Torre, V. "Ill-Posed Problems and Variational Principles in Vision", *A. I. Memo 773*, MIT, 1984.

Poggio, T., Voorhees, H. and Yuille, A. "A regularized solution to edge detection", in preparation, 1984

Pogorelov, A. V. *Differential geometry.* P. Moozdhoff N. V. Groningen, 1965.

Poston, T. and Stewart, I. N. *Taylor expansions and catastrophes.* Pitman Publishing, London, 1976.

Rosenfeld, A. "Quadtrees and Pyramids: hierarchical representation of images" TR 1171, University of Maryland, 1982

Rosenfeld, A. and Kak, A. C. *Digital Picture Processing*, second edition, Academic Press, New York, 1982.

Schoenberg, I. J., "Contributions to the problem of approximation of equidistant data by analytic functions." *Quart. Appl. Math.*, **4**, 45-99, 112-141, 1946.

Schoenberg, I. J., "Spline functions and the problem of graduation." *Proc. nat. Acad. Sci. USA*, **52**, 947-950, 1964.

Shanmugam, K. F., Dickey, F. M. and Green, J. A. "An optimal frequency domain filter for edge detection in digital pictures." *IEEE Trans. Pattern Anal. & Mach. Intell.*, **PAMI-1**, 37-49, 1979.

Slepian, D. "Some asymptotic expansions for prolate spheroidal wave functions." *J. Math & Phys.*, **44**, 99-149, 1965.

Stansfield, J. L. "Conclusions from the commodity expert project", *A. I. Memo 601*, MIT, 1980.

Thom, R. "Quelques proprietes globales des varietes differentiables." *Comm. Math. Helv.*, **28**, 17-86, 1954.

Tikhonov, A. N. "Regularization of incorrectly-posed problems." *Soviet Math. Dokl.*, **4**, 1624-1627, 1963.

Tikhonov, A. N., Arsenin, V. Y. *Solution of Ill-Posed Problems*, Winston and Wiley Publishers, Washington, 1977.

Widder, D.V. *The heat equation* , Acad. Press, New York, 1975.

Witkin, A. "Scale-Space Filtering", Proceedings of IJCAI, 1019-1021, Karlsruhe, 1983.

Yuille, A.L. and Poggio, T. "Scaling Theorems for Zero-crossings". *A. I. Memo 722*, MIT, 1983a.

Yuille, Poggio, "Fingerprints Theorems for zero-crossings", *A. I. Memo 730*, MIT, 1983b.

1. Appendix: Differentiation through Taylor expansion

In previous sections, we have seen that to safely perform differentiation, it is necessary to smooth the data by some appropriate (analog or digital) filtering. If this filtering has removed enough high frequencies, so that our filtered image is band-limited function, and by the Paley-Wiener theorem (Boas, 1954) is also an entire function (see Section 3), numerical differentiation can be performed in a computationally more efficient way through Taylor expansion.

If $f(x)$ is entire, than $f(x)$ is also analytic and the Taylor series

$$f(x) = f_k + (x - x_k)f'_k + \ldots + \frac{(x - x_k)}{n!}f_k^{(n)} + \ldots \qquad [1]$$

has an infinite radius of convergence. If we have $2n + 1$ sampled points from [2], we can obtain $2n$ linear equations from which we can solve for $f_k^{(j)}$, $j = 1, 2, \ldots, 2n$.

Three equidistant points give

$$\begin{aligned} f'_k &= \frac{1}{2n}(f_{k+1} - f_{k-1}) \\ f''_k &= \frac{1}{h^2}(f_{k-1} - 2f_k + f_{k+1}) \end{aligned} \qquad [3]$$

With five equidistant points, we obtain

$$\begin{aligned} f'_k &= \frac{1}{12h}(f_{k-2} - 8f_{k-1} + 8f_{k+} - f_{k-2}) \\ f''_k &= \frac{1}{12h^2}(-f_{k-2} + 16f_{k-1} - 30f_k + 16f_{k+1} - f_{k+2}) \end{aligned} \qquad [4]$$

When the performances of the numerical differentiation obtained through spline interpolation (eqs. 2.9 - 2.10) are compared with those obtained by Taylor expansion (equations 3-4 in this appendix), it turns out that the first method gives more accurate and consistent results with noisy data while the second method is more efficient with data that are already smooth.

2. Appendix: Sampling

Since image processing is performed in terms of discrete representations of signals and filters, it is important that manipulations of sampled images and filters have a meaningful connection with the original image $I(x)$ and the analytic form of the filter $f(x)$. More precisely, for linear filtering, if I_i is a discrete sequence of points of $I(x)$, and f_i is another discrete sequence of points of $f(x)$, the discrete convolution

$$g_i = \sum_k I_k \cdot f_{k-i} \qquad [1]$$

should be related to the exact convolution

$$g(x) = \int I(y) f(y - x) dy = I(x) * f(x). \qquad [2]$$

This relation is clarified by standard results (Oppenheim and Johnson, 1972; Borsellino and Poggio, 1973).

Suppose that we may represent $I(x)$ and $f(x)$ as

$$I(x) = \sum_i I_i \phi_i(x) \qquad [3a]$$

$$f(x) = \sum_i f_i \phi_i(x) \qquad [3b]$$

where $\phi_i(x) = \sin c[\frac{\pi}{h}(x - ih)]$. Then

$$I(x) * f(x) = g(x) = \sum_i g_i \phi_i(x), \qquad [4]$$

where $g_i = \sum_k I_k f_{k-i}$.

Thus from the discrete convolution of the sampled values [2], it is possible to recover completely, the exact convolution of the original image with the filter. It is now possible to represent a signal in the form of equation [1] when the signal is band-limited and correctly sampled. If one uses band-limited filters it is possible to obtain the required representation [3b] for the filter. For an arbitrarily sampled image, however, it is difficult to obtain the required representation [3a]. Indeed it would be necessary to sample the image according to the cutoff frequency of the optical system used in the imaging process, which is generally too high to be of practical use. This is related to the classical problem of aliasing. The simplest way to obtain a reasonable solution to the problem is to initially filter the image with an appropriate band-limited filter, before any further operation.

3. Appendix: The geodesic curvature

It may be of some interest to ask whether the line on the surface $X = (x, y, f(x, y))$, whose tangent is always in the direction of the gradient is also a geodesic. A geodesic is aline whose geodesic curvature K_g is always zero. In what follows we will briefly answer this question The surface X has the first fundamental form [4.6].

As shown later, the geodesic curvature of a curve on surface X whose tangent is always in the direction of the gradient is

$$K_g = \frac{g\left[\frac{f_{xy}}{f_x}\left(\frac{f_y^2}{f_x^2} - 1\right) + \frac{f_y}{f_x^2}(f_{yy} - f_{xx})\right]}{\left\{g + \frac{1}{f_x^2}\left[1 + f_y^2(f_x^2 + f_y^2)\right]\right\}^{3/2}}. \qquad [1]$$

Now, $K_g = 0$, when $f_x^2 = f_y^2$ and $f_{xx} = f_{yy}$, that is, for surface of revolution. Therefore the line on surface X, which is always directed along the gradient, has the following properties:

(1) its normal curvature K_n is equal to $\frac{1}{g^3} \times \frac{\partial^2(f)}{\partial n^2}$.

(2) it is a geodesic, only for surface of resolution.

Let us now compute K_g. Given that the coefficients of the first fundamental form are

$$E = 1 + f_x^2 \quad F = f_x f_y \quad G = 1 + f_y^2,$$

The associated Cristoffel symbols $\Gamma_{11}^1 \quad \Gamma_{11}^2$ satisfy the equations

$$\Gamma_{11}^1 EF + \Gamma_{11}^2 F^2 = \frac{1}{2} F E_x \qquad [2]$$

$$\Gamma_{11}^1 EF + \Gamma_{11}^2 EG = F_x E - \frac{1}{2} E E_y. \qquad [3]$$

or

$$\Gamma_{11}^2 = \frac{\frac{1}{2}(F E_x + E E_y) - F_x E}{F^2 - EG} \qquad [4]$$

$$\Gamma_{11}^1 = \frac{1}{2}\frac{E_x}{E} - \frac{\frac{1}{2}\left(F\frac{E_x}{E} - E_y\right) - F_x}{F - \frac{EG}{F}}. \qquad [5]$$

Finally,

$$\begin{aligned} \Gamma_{11}^2 &= \frac{\frac{1}{2}(f_x f_y 2 f_x f_{xx} + (1 + f_x^2) 2 f_x f_{xy}) - (1 + f_x^2)(f_{xx} f_y + f_x f_{xy})}{f_x^2 f_y^2 - 1 - f_x^2 - f_y^2 - f_x^2 f_y^2} \\ &= \frac{f_x^2 f_y f_{xx} - (1 + f_x^2) f_{xx} f_y}{-g^2} = \frac{f_{xx} f_y}{g^2} \end{aligned} \qquad [6]$$

$$\begin{aligned} \Gamma_{11}^1 &= \frac{1}{2}\frac{E_x}{E} - \Gamma_{11}^2 \frac{F}{E} = \frac{f_x f_{xx}}{1 + f_x^2} - \frac{f_{xx} f_y}{g^2}\frac{f_x f_y}{1 + f_x^2} \\ &= \frac{f_x f_{xx}(1 + f_x^2 + f_y^2) - f_{xx} f_y^2 f_x}{(1 + f_x^2) g^2} \\ &= \frac{f_x f_{xx}}{g^2}. \end{aligned} \qquad [7]$$

Similarly, Γ^1_{12} Γ^2_{12} satisfy the equations

$$\Gamma^1_{12} EF + \Gamma^2_{12} F^2 = \frac{1}{2} F E_y \qquad [8]$$

$$\Gamma^1_{12} EF + \Gamma^2_{12} EG = \frac{1}{2} E G_x (4). \qquad [9]$$

Some algebra gives from [9] and [10]

$$\Gamma^1_{12} = \frac{f_x f_{xy}}{g^2}. \qquad [10]$$

$$\Gamma^2_{12} = \frac{f_y f_{xy}}{g^2}. \qquad [11]$$

Now suppose that $X = (x, y, f(x, y))$ is a regular parametrization of our surface around P, and let $x = x(t), y = y(t)$ be the equation of the curve γ in a neighborhood of P. Then (Pogorelov, 1965)

$$K_g = \frac{\sqrt{EG - F^2}}{\left(Ex'^2 + 2Fx'y' + Gy'^2\right)^{3/2}} \{x''y' - y''x' + Ay' - Bx'\} \qquad [12]$$

is the geodesic curvature of γ in P, where

$$A = \Gamma^1_{11} x'^2 + 2\Gamma^1_{12} x'y' + \Gamma^1_{22} y'^2 (6a) \qquad [13a]$$

$$B = \Gamma^2_{11} x'^2 + 2\Gamma^2_{12} x'y' + \Gamma^2_{12} y'^2. \qquad [13b]$$

If we want to compute the geodesic curvature of a curve γ which lies on the surface X and always has the tangent vector in the direction of the gradient of $f(x, y)$, we must find the equation of the curve $(x(t), y(t))$. Let us suppose

$$x(t) = t \qquad [14]$$

and therefore

$$\frac{dy}{dt} = \frac{f_y(x(t), y(t))}{f_x(x(t), y(t))}. \qquad [15]$$

Now equation [15] defines a differential equation in $y = y(t)$ whose solution completes the equation of the curve. We have $x(t) = x$ and $y(t)$ such that

$$x'(t) = 1 \qquad x''(t) = 0$$

$$y'(t) = \frac{f_x}{f_y} \qquad y''(t) = \frac{f_{xy}}{f_x}\left(\frac{1 - f_y^2}{f_x^2}\right) + \frac{f_y}{f_x^2}(f_{yy} - f_{xx}). \qquad [16]$$

Now let us compute

$$
\begin{aligned}
x''y' - y''x' + Ay' - Bx' &= -\frac{f_{xy}}{f_x}\left(1 - \frac{f_y^2}{f_x^2}\right) + \frac{f_y}{f_x^2}(f_{yy} - f_{xx}) \\
&\quad + \frac{f_y}{g^2 f_x}\left[f_x f_{xx} + 2 f_x f_{xy}\frac{f_y}{f_x} + f_x f_{yy}\frac{f_y^2}{f_x^2}\right] \\
&\quad - \frac{1}{g^2}\left[f_y f_{xx} + 2 f_y f_{xy}\frac{f_y}{f_x} + f_y f_{yy}\frac{f_y^2}{f_x^2}\right] \\
&= \frac{f_{xy}}{f_x}\left(\frac{f_y^2}{f_x^2} - 1\right) + \frac{f_y}{f_x^2}(f_{yy} - f_{xx}) + \frac{1}{g^2}\{0\}.
\end{aligned}
\tag{17}
$$

Now

$$
\sqrt{EG - F^2} = g
$$

and

$$
\begin{aligned}
Ex'^2 + 2Fx'y' + G y'^2 &= \left(1 + f_x^2\right) + 2 f_x f_y \frac{f_y}{f_x} + \left(1 + f_y^2\right)\frac{f_y^2}{f_x} \\
&= 1 + f_x^2 + 2 f_y^2 + \frac{1}{f_x^2} + \frac{f_y^4}{f_x^2} \\
&= \frac{1}{f_x^2}\{f_x^2 + f_x^4 + 2 f_x^2 f_t^2 + 1 + f_y^4\} \\
&= \frac{1}{f_x^2}\{f_x^2\left(1 + f_y^2\right) + f_x^4 + 1 f_y^2\left(f_x^2 + f_y^2\right) + 1\}.
\end{aligned}
\tag{18}
$$

Finally, we have

$$
\begin{aligned}
K_g &= \frac{g \cdot \left[\frac{f_{xy}}{f_x}\left(\frac{f_y^2}{f_x^2} - 1\right) + \frac{f_y}{f_x^2}(f_{yy} - f_{xx})\right]}{\{\frac{1}{f_x^2}\{f_x^2\left(1 + f_x^2 + f_y^2\right) + 1 + f_x^2 f_y^2 + f_y^4\}\}^{3/2}} \\
&= \frac{g\left[\frac{f_{xy}}{f_x}\left(\frac{f_y^2}{f_x^2} - 1\right) + \frac{f_y}{f_x^2}(f_{yy} - f_{xx})\right]}{\{g + \frac{1}{f_x^2}\{1 + f_y^2\left(f_x^2 + f_y^2\right)\}^{3/2}\}}.
\end{aligned}
\tag{19}
$$

The geodesic curvature K_g is zero when

$$
\frac{f_{xy}}{f_x}\left(\frac{f_y^2}{f_x^2} - 1\right) + \frac{f_y}{f_x^2}(f_{yy} - f_{xx}) = 0. \tag{20}
$$

Eq. [20] is not generally satisfied; it is satisfied for surfaces such that

$$
\begin{aligned}
f_x^2 &= f_y^2 \\
f_{yy} &= f_{xx}
\end{aligned}
\tag{14}
$$

as a sphere or a surface of revolution.

4. Appendix: Approximation and interpolation

The traditional procedure for performing numerical differentiation of sampled functions is to interpolate using polynomials or related functions and to analytically differentiate the interpolated function. This procedure is usually justified if the interpolated function converges uniformly to the original function as the number of samples increases. Most of the classical results in the interpolation and approximation of functions deal with this problem. In the case of images the main problem is however different: approximation for the purpose of differentiation has to be robust against noise. The solution to this problem has to be sought in regularization theory, as we explained earlier. In this appendix, we discuss some of the classical results on interpolation and approximation for completeness and not because they are directly relevant to the problem of regularizing edged detection. The reader will notice, however, that there are several connections between the classical results outlined here and our approach described in the the main body of this paper.

Uniform convergence for polynomial interpolation is not guaranteed even when the original function is C^∞, or when it is analytical. Uniform convergence on bounded sets requires the original function to be *entire*. From the Paley-Wiener theorem (Boas, 1954) we know that this is the case with band-limited functions.

If the original function is analytic, numerical differentiation may be performed using an appropriate Taylor expansion without interpolation. Interestingly, almost every function is made analytic by filtering with a Gaussian, since the filtered function is a solution of the heat equation (Widder, 1975). Thus, in order to safely perform numerical differentiation, it is necessary to use band-limited or Gaussian filters.

Note that if approximation (in the Weierstrass-Bernstein sense) rather than interpolation is used, we obtain uniform convergence for all continuous bounded functions on bounded sets. Therefore, differentiation through approximation is successful on bounded sets for all bounded C^1 functions. We will see, however, that convergence is too slow for this approach to be practical.

In what follows, we will use a one-dimensional approach for the sake of simplicity, but all our conclusions and results can be easily extended to two dimensions.

4.1. *Interpolation and differentiation*

Consider a function $f(x)$ defined in $[a, b]$ and have $a \leq x_o < x_1 < x_2 < \ldots < x_k < \ldots < x_n \leq b$ distinct points, and

$$f_k = f(x_k) \qquad [1]$$

the values of f at x_k. It is well known that there exists a polynomial $p_n(x)$ of degree n such that for the given values $x_o < x_1 \ldots < x_n$ takes the values $y_0, y_1 \ldots, y_n$ (Davis, 1975). This polynomial is

$$p_n(x) = \sum_0^n y_k l_k(x), \qquad [2]$$

where $l_x(x)$ are Lagrange polynomials

$$l_k(x) = \frac{(x - x_o)(x - x_1)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n)}{(x_k - x_o)(x_k - x_1)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n)}. \quad [3]$$

From equation [2] and [3], we consider the best way to estimate the derivative of $f(x)$ in x_k, $f'(x_k)$ by computing

$$p'_n(x_k) = \sum_0^n y_k l'_k(x) \quad _{x=x_k}. \quad [4]$$

In order to use this procedure reliably we need to know that in some way $p_n(x)$ is a good approximation of $f(x)$ outside the sampled points. We would like to know that by increasing the number n of sampled points x_n in $[a, b]$, we have

$$\lim_{n\to\infty} \quad p_n(x) = f(x). \quad [5]$$

This is equivalent to uniform convergence. At the beginning of the century, Bernstein (see P. Davis, 1975) proved that equidistant interpolation over $|x| \leq 1$ to the function $y = |x|$ diverges for $0 < |x| < 1$; that is, continuity of $f(x)$ is not sufficient to ensure uniform convergence. Runge showed that even if $f(x)$ is analytical in $[a, b]$ uniform convergence may fail (P. Davis, 1976). For the function $f(x) = 1/(1 + x^2)$, Runge showed that if $p_n(x)$ interpolates $f(x)$ at equidistant points in $[-5, 5]$, $p_n(x)$ converges to f only in $|x| < 3.63\ldots$ and diverges outside the interval. Although $f(x)$ is analytic in $\Re$ is not analytic in the complex plane C and the singular points $\pm i$ induce this divergence (see P. Davis, 1975). To obtain uniform convergence of $p_n(x)$ to $f(x)$ in $[a, b]$, it is necessary for $f(z)$ to be analytic in a subregion of the complex plane C containing the segment $[a, b]$ (see Theorem 4.3.1 of P. Davis, 1975).

It may be useful to remember that polynomial interpolation is not optimal if the sampled points are equidistant. Polynomial interpolation is optimal if interpolation of $f(x)$ in $[0, 1]$ of order n is carried at the zeros of the Chebychev polynomials $T_n(x)$. In general, the procedure of differentiation through interpolation with polynomial or related functions may badly fail when applied to arbitrary sampled functions.

4.2. *Differentiation of analytic-, entire- and band-limited functions*

If we know that $f(x)$ is analytic in $[a, b]$ and we have no information about its behavior on the complex plane C we cannot safely use differentiation through interpolation. If $f(x)$ is analytic in $x_k \in [a, b]$, we have

$$f(x) = f(x_k) + (x - x_k)f^{(1)}(x_k) + \ldots + \frac{(x - x_k)^n}{n!} f^{(n)}(x_k) + \ldots \quad [6]$$

in $[x_k - \delta, x_k + \delta]$. If we have $2n + 1$ sampled points in $[x_k - \delta, x_k + \delta]$, from equation [6], we can obtain $2n$ linear equations, from which we can solve for $f^{(j)}(x_k), j = 1\ldots 2n$. In the case of three or five equidistant points we obtain the formulae shown in Appendix 1.

The main problem with this procedure of numerical differentiation is that we do not generally know the radius of convergence of the Taylor expansion [6] and, given an arbitrary sampled analytical function, we do not know how many points around x_k fall in the convergence interval. When the class of original functions f is further restricted to entire functions, the

Taylor expansion [6] is valid in $\Re$, that is, it has an infinite radius of convergence, and the above procedure can be carried out safely.

Now let us suppose that $f(x)$ is entire, $f \epsilon L^2(\Re)$, and

$$\lim_{\rho \to \infty} \sup \frac{\ln M(\rho)}{\rho} = \delta \qquad [7]$$

where

$$M(\rho) = \max_{|z|=\rho} |f(z)| \qquad [8]$$

with $z \in C$ and f extended to the complex plane.

If $\delta < \infty$ then $f(x)$ is said to be of exponential type δ and by the Paley-Wiener Theorem (Boas, 1954; Achieser, 1956), $f(x)$ is band limited with $F(\omega) = 0$ for $|\omega| \geq \delta$, where $F(\omega)$ is the Fourier transform of $f(x)$. The Paley-Wiener Theorem represents the connection between entire and band limited functions. It may be useful to remember that the Gaussian is entire and belongs to $L^2(\Re)$ but falls off just too quickly to be of finite exponential type and therefore to be band-limited.

If $f(x)$ is band-limited (with cutoff frequency ω_o), and $f(x)$ has been correctly sampled at equidistant points spaced h, the Shannon Sampling Theorem gives us

$$f(x) = \sum_{-\infty_i}^{+\infty} f_i \operatorname{sinc}\left[\frac{\pi}{h}(x - ih)\right], \qquad [9]$$

where $\operatorname{sinc} x = \frac{\sin x}{x}$. In this case, we can compute f'_k exactly from [9] as

$$\begin{aligned} f'_k &= \frac{1}{h} \sum_{-\infty_{i \neq k}}^{+\infty} f_i \frac{\cos \pi(k-i)}{k-i} \\ &= \frac{1}{h}\left[(f_{k+1} - f_{k-1}) - \frac{1}{2}(f_{k+2} - f_{k-2}) + \frac{1}{3}(f_{k+3} - f_{k-3}) - \ldots\right]. \end{aligned} \qquad [10 <]$$

Although [10] is exact for correctly sampled band-limited functions, it converges rather slowly.

4.3. *Approximation and Differentiation*

It is well known that if $f(x)$ is continuous in $[a, b]$, for every given $\epsilon > 0$, we can find a polynomial $p_n(x)$ of sufficiently high degree such that

$$|f(x) - p_n(x)| \leq \epsilon \quad a \leq x \leq b. \qquad [18]$$

This is the so-called Weierstrass approximation theorem. Now let us suppose that we have an $f(x)$ continuous and bounded in $[0, 1]$ and we know the values of $f(x)$ at equidistant points $x_k = \frac{k}{n}$. Instead of interpolating a polynomial through the known points, we can construct the Bernstein polynomial:

$$B_n(x) = \sum_{k=0}^{n} f(x_k)\binom{n}{k}x^k(1-x)^{n-k}. \quad [11]$$

Observe that $B_n(0) = f(0)$ and $B_n(1) = f(1)$, but apart from 0 and 1 $B_n(x)$ is not in general equal to $f(x_k)$. A fundamental theorem of Berstein shows that

$$\lim_{n\to\infty} B_n(x) = f(x) \quad \text{in} \quad [0,1]. \quad [13]$$

Moreover, if $f(x) \in C^1$, we have

$$\lim_{n\to\infty} B_n'(x) = f'(x) \quad \text{in} \quad [0,1]. \quad [14]$$

Thus, Bernstein polynomials provide simultaneous approximation of the function and its derivatives. If we want to obtain an estimation of the derivative of a sampled function, we can construct the Bernstein polynomial from the sampled values, compute the analytical derivative of $B_n(x)$ and take it as an estimate of $f'(x)$. The drawback of Bernstein polynomials is that their convergence is very slow as shown by the poor performance of a 3- or 5- point approximation. Therefore, if we have samples of a generic function of class C^1, we can use Bernstein polynomials to obtain an estimation of the values of derivatives at sampled points, but many points are needed for a good estimate. If we have samples of an entire function, the most efficient procedure is to use the equations derived in Appendix 1.